White Paper

Report ID: 109479

Application Number: HG-50041-13

Project Director: Sheldon Pollock

Institution: Columbia University

Reporting Period: 5/1/2013-4/30/2017

Report Due: 7/31/2017

Date Submitted: 7/17/2017

SARIT: Enriching Digital Collections in Indology

NEH-DFG Bilateral Humanities Program, Grant No. HG-5004113

Final Report (July 2017)

Andrew Ollett and Sheldon Pollock

1. Introduction and Background

This is the final report of the "Columbia subproject" of the four-year project called "SARIT: Enriching Digital Collections in Indology," which was the recipient of an NEH-DFG Bilateral Digital Humanities Grant in 2013. The principal investigator on this project was Sheldon Pollock, Arvind Raghunathan Professor of South Asian Studies at Columbia University. Andrew Ollett also worked on the project, first as a graduate student assistant (2013–2015) and subsequently as a postdoctoral affiliate (2015–2017). Our partners in this bilateral project were the University of Heidelberg. The principal investigator of the "Heidelberg subproject" was Birgit Kellner, Professor of Buddhist Studies. This report focuses on the experience of the Columbia subproject, but reflects the results that both subprojects achieved by working together over the course of the funding period.

BACKGROUND. SARIT was originally conceived by two Sanskrit scholars, Dominik Wujastyk and Richard Mahoney, in 2008. By this time, textual scholars in other fields, such as Classics and East Asian Studies, had created collections of digital texts that were widely used in actual research. By contrast, no such resources were available to scholars of South Asian Studies. Collections of digital texts, in Sanskrit and other South Asian languages, were assembled haphazardly, in a wide range of formats, and in varying degrees of quality. There was no resource comparable to the *Thesaurus Linguae Graecae* or the *Perseus Project* in Classics that could offer access to digital texts in a standard format and of a reliably high quality. Wujastyk and Mahoney imagined that adhering to the guidelines of the Text Encoding Initiative (TEI), which were increasingly used for digital-texts projects in other fields, might address some of these problems. In principle, TEI documents offered three advantages over "plain-text" formats: (1) it offered a unified approach to encoding various aspects of the text, from its structure (prose paragraphs or verse lines) to "named entities" like people and places, to textual variation; (2) it provided a uniform way of reporting the *source* or *sources* on which the digital text was based in the document header, thus avoiding the widespread problem of users not knowing which printed editions a particular digital text was based on; (3) it could be continuously improved, as users added new elements of markup, or corrected typos, and those improvements could also be explicitly documented in a "revision description" of the document header. In a word, TEI offered a principled approach to document *metadata*, information about how the document was prepared and encoded, what sources it was based on, and when and how it was revised. Wujastyk and Mahoney added TEI markup to a number of already-existing digital texts in

Sanskrit and created a website, using the University of Chicago's Philologic software, where users could search and browse these TEI documents. They named this project SARIT, the Sanskrit word for "river," which stands for ***s**earch **a**nd **r**etrieval of **I**ndic **t**exts*.

In order to realize the potential of this new approach to providing digital texts in Sanskrit, however, the SARIT project needed to expand in three main areas. First, it had to increase the number of digital texts on offer. Second, it had to improve the web interface through which users accessed these texts. Third, and most importantly, it had to come up with standards and practices for the incipient field of South Asian digital philology: everything, from the way digital texts were produced, to the standards followed in producing them, to the software used to interact with them, had to be envisioned more or less from scratch, given the absence of models in the field that could be adapted and extended. Indeed, there were—and remain—real questions about the potential of digital philology for South Asian Studies, about what it actually involved, and how if at all it differed from traditional approaches in its methods, or in its questions, or in its results. To better address these technical, organizational, and scholarly questions, SARIT formed an advisory board in 2012, which included Sheldon Pollock and Birgit Kellner. One of the first priorities of the board was to seek funding for SARIT, and this led to the successful application for a Bilateral Digital Humanities grant from the NEH and DFG in 2013.

This new model led to some changes in the way that SARIT, as a project and as an organization, was conceived. Mahoney left the SARIT project, and now runs an independent project called *Indica et Buddhica*. The development of SARIT would now take place under a number of different institutions, in the context of specific funded projects. From 2013 to 2017, the development took place at Columbia University and the University of Heidelberg, under the direction of Sheldon Pollock and Birgit Kellner respectively. This project was planned as an "enhancement" to SARIT as it existed at the time, and as a substantial step toward realizing the longer-term vision of SARIT as a sustainable collection of high-quality digital texts produced and maintained by the scholarly community. Thus, for the past several years, we have operated with the vision of a "big SARIT" (*mahāsarit*) that exists independent of any institution or funding sources, representing the collection of texts itself and the tools for interacting with them, alongside a number of "tributary SARITs" (*avāntarasarit*) that flow into it, enhancing the overall project with contributions of texts and software development.

The NEH-DFG grant was the largest that SARIT had received, and marked its transformation from a pilot project to a recognized presence in South Asian Studies. Before 2013, SARIT had largely been maintained by Wujastyk, with the assistance of Patrick McAllister, a colleague at the University of Vienna. Now, in 2017, SARIT is a truly international undertaking. The collection of texts features contributions from all over the world; the web application represents the collaboration of project staff in the United States and Germany with software developers in Germany; and both texts and the software are made available for free, to anyone in the world, under a Creative Commons license. Besides the individual goals they set out to accomplish within the terms of the funded project, the Columbia and Heidelberg subprojects therefore had to contend with

a major shift in the scale of the project, and the organizational and technical challenges that this shift implies.

This document will review what we, the Columbia subproject, accomplished under the four years of NEH funding for SARIT. We compare these accomplishments to the goals that we announced at the start of the project, and we reflect on some of the persistent challenges that we've faced. We then look to the future. What remains to be done for SARIT, and for digital resources in South Asian Studies more generally? How do we imagine the "ecosystem" of digital resources in the field will look five, ten, or fifteen years from now? What will the community of users look like, and what kinds of things will they want to do with those resources? What lessons can we learn from the past four years? What should we do differently in the future? What are the most promising avenues of future work?

2. Achievements and Reflections

Work began on the NEH-DFG project in the summer of 2013. The Columbia and Heidelberg subprojects took up different aspects of the work according to their different institutional resources and intellectual priorities. For the preparation of digital texts, the Columbia subproject would focus on poetics (*alaṃkāraśāstra*) and hermeneutics (*mīmāṃsā*), while the Heidelberg subproject would focus on epistemology (*pramāṇavidyā*). Both subprojects were to collaborate in the formulation of specific guidelines for adding TEI encoding to Sanskrit texts. The Heidelberg subproject was to lead the development of a new web application for SARIT, given the opportunity to work with in-house developers at the Heidelberg Research Architecture (HRA). For its part, the Columbia subproject was to revive a defunct bio-bibliographic database associated with the "Sanskrit Knowledge Systems on the Eve of Colonialism" (SKSEC) project, and consider the possibility of integrating SARIT's collection of digital texts with this new bio-bibliographic resource. After summarizing the organization and administration of the Columbia subproject, this review will consider four areas of work in turn: the production of digital Sanskrit texts, the development of a web-based interface, the development of a bio-bibliographical database, and the promotion of standards and best practices for the digital philology of South Asia.

PERSONNEL. Sheldon Pollock and Andrew Ollett have worked on the project over the duration of the funding period. In addition, the Columbia subproject has hired a number of shorter-term undergraduate, graduate and postdoctoral assistants. All of these assistants worked on the preparation of TEI documents from the "raw" texts provided by our double-keyboarding vendor in India. In 2013–2014, Ollett trained two graduate students, Shiv Subrahaniam and Jay Ramesh, to do this encoding work. In 2014–2015, Ollett again trained two undergraduate students, Gayathri S and Sireesh Gururaja, to do this work. Finally, in 2015–2016, Columbia hired a postdoctoral researcher, Dheepa Sundaram, to assist with the preparation of TEI documents. None of these assistants had any experience in TEI, and by and large, the addition of TEI encoding to the text documents proceeded very slowly. Gururaja, however, had some experience in computer programming, and wrote a number of scripts that automated some of the most important text-encoding tasks. In the

early stages of the project, this task was also complicated by the lack of explicit guidelines for encoding certain features of Sanskrit texts. By 2015, however, the Columbia and Heidelberg subprojects had produced a draft of encoding guidelines that were specifically tailored to Sanskrit texts (see section 2D below), and our postdoctoral assistant Dheepa Sundaram referred to these guidelines while preparing the TEI documents.

ORGANIZATION. At the commencement of the NEH-DFG grant, both subprojects began to use Redmine, an open-source task management application, to keep track of project-related tasks. We have continued to use Redmine for the duration of the grant. We found it to be effective overall, despite some minor bugs that made it difficult to search for issues. In addition, we held regular meetings over Skype or Google hangouts to discuss progress on outstanding issues. Towards the beginning of the project, these were monthly meetings. Although we have often discussed issues with text encoding, the primary purpose of these monthly meetings has been to check in with the software developers. Due to staffing changes at the Heidelberg Research Architecture, development on the web platform slowed in 2015–2016 (see below), and as a consequence these meetings also became less frequent. In 2016–2017, we have returned to the monthly schedule for project meetings.

The data produced by the project can be classified into text data and application data. The text data represents the texts that the Columbia and Heidelberg subprojects have prepared for SARIT. The application data represents the source code of the web application. Each subproject had its own way of keeping track of the text data while it was being prepared. At the commencement of the project, each subproject maintained a separate Dropbox folder containing the text data to which all of the subproject staff had access. In 2015, the Columbia subproject replaced its Dropbox folder with a private GitHub repository, which offers a more robust approach to version control. When either subproject finalized a text, the staff would deposit the text in a public GitHub repository for the entire project. This is the "canonical" repository for the SARIT project, and the SARIT web application is populated with data from this repository. The application data is also maintained in a public GitHub repository, although it has moved several times over the course of the project as it has undergone different stages of development by different groups of people.

COSTS. Given that the Heidelberg subproject contracted the majority of the software development, the Columbia project's costs fall under two main categories: (1) payments to our double-keyboarding vendor in India (approximately 20% of our total costs); and (2) compensation for the subproject staff (approximately 75% of our total costs). In addition, we made several payments to a software developer in Israel to convert the data from the erstwhile SKSEC database. See the attached budget for details.

CONFERENCES AND PRESENTATIONS. Ollett presented ongoing work on SARIT at four scholarly conferences over the duration of the project. The first was at the University of Oxford in September 2014, at a conference on the theme of "Buddhism and the Digital Humanities" organized by Jan Westerhoff. That presentation dealt with the plans to add functionality to SARIT beyond simply searching through Sanskrit texts, and in particular, the possibility of using SARIT as a platform for collaboratively editing Sanskrit texts, as was originally envisioned in our proposal to

the NEH.[1] The second was at the University of British Columbia, in Vancouver, in January of 2016, at a conference on the the theme of "Digital Textualities of South Asia" organized by Adheesh Sathaye. That presentation dealt with new possibilities for structuring the text data in the SARIT corpus, and in particular the use of canonical reference systems and linked open data.[2] There have also been two meetings specifically organized around SARIT. Both have taken place at the Institute for the Cultural and Intellectual History of Asia, part of the Austrian Academy of Sciences in Vienna, where one of the project's principal investigators (Kellner) has been director since 2015. In June 2016, there was a one-day meeting in Vienna where most of the people involved with SARIT, including Ollett, presented their work. In May 2017, there was a three-day meeting in Vienna, which was devoted to SARIT's present and future role in the "digital ecosystem" of South Asian Studies, titled "The Future of Digital Texts in South Asian Studies." This summative workshop included participants from many of the other digital projects in South Asian Studies (see the attached program and abstracts). Ollett presented a retrospective on the Columbia subproject's work, and a prospective vision of how digital texts must be provisioned in the future.[3] The Heidelberg subproject presented SARIT at a number of other venues, including the meeting of the International Association of Buddhist Studies in Vienna in the summer of 2015. In addition, members of the SARIT project held a number of informal meetings at scholarly conferences over the past four years, including the World Sanskrit Conference in Bangkok in the summer of 2015. (See section 4, "Bibliography and References," for a complete list of presentations of SARIT at scholarly conferences.)

As an introduction to the achievements of the SARIT project, we can refer to the results of a survey organized by one of SARIT's founders, Dominik Wujastyk, and presented by him at the aforementioned "SARIT Workshop" in Vienna in May 2017.[4] Although only twenty people completed the survey, which was posted on the INDOLOGY mailing list, they were largely unanimous in their appreciation of the quality, usefulness, and openness of SARIT. Since "quality," in the sense of carefully typed-in texts with a high degree of textual markup, was one of the principal goals of the project, it is important to note that all twenty respondents either found the quality of SARIT to be "high" (9) or "very high" (11). Most of the respondents suggested increasing SARIT's coverage. "More texts," as one respondent said; others suggested that SARIT provide texts that have a wider readership, relative, of course, to the corpus of Sanskrit literature. By and large, however, scholars are interested in *using* SARIT, and find the texts provided on the SARIT website useful and easy enough to interact with, and the overarching goal for the future will be to turn SARIT from a "proof-of-concept" for digital Sanskrit texts into an essential and general resource for scholars of South Asian Studies.

1 https://www.academia.edu/8255307/Sarit-pras%C4%81ra%E1%B9%87am_Developing_SARIT_beyond_Search_and_Retrieval.
2 https://prezi.com/-mde_4yynp_d/treasury-department-anthologies-in-the-digital-age/
3 https://www.academia.edu/33883303/A_Less_Distant_Future_Sanskrit_Texts_for_Scholarly_Communities_in_the_Digital_Age.
4 https://www.academia.edu/33153494/What_Do_Users_Want_From_SARIT_in_Future

A. Production of Digital Sanskrit Texts

The primary task that the Columbia subproject had set for itself was the production of high-quality digital Sanskrit texts that would be made available to the public through SARIT. As noted above, this subproject planned to contribute texts that belonged to two traditions of systematic thought: poetics (*alaṃkāraśāstra*) and hermeneutics (*mīmāṃsā*). Since OCR technology for Devanāgarī, the script in which most Sanskrit texts are printed, was not up to a sufficient standard of accuracy at the beginning of the project, we decided that we would have the texts "double-keyboarded" by a firm in India. This means that two people independently type up the text, and the resulting documents are compared against each other and harmonized. Various firms promise a high degree of accuracy in double-keyboarding, between 99% and 99.5%, although the accuracy depends on the quality of the scanned documents that are provided to the firm. Throughout the project, we have used SWIFT Information Technologies in Mumbai as our double-keyboarding vendor. They have delivered reasonably accurate files at competitive rates and, importantly, with a very fast turnaround time. It has, however, proven necessary to check their work closely and hold them to their promises, since they have occasionally sent the wrong files by mistake, or misread or ignored the instructions we provided to them. After we receive the double-keyboarded files, we then convert them into the target format: TEI documents that are encoded with a high level of detail and a high degree of consistency across the SARIT corpus. This "post-processing" phase is the most time-consuming and labor-intensive, since it usually requires someone who is familiar with the structure and format of the original printed editions—therefore someone who knows Sanskrit and knows the editorial conventions of printed Sanskrit texts—to add structural and logical markup to the text files, for example dividing them into chapters or section, or numbering verses, or reformatting footnotes. As explained below, many of these tasks had to be performed manually. At the end of the process, the entire file is checked against a schema to ensure that the TEI guidelines have been followed correctly, and then it is uploaded onto the public SARIT repository. It then can be accessed there, as a raw XML file, or alternatively it can be accessed from the SARIT web application.

Although we have maintained this workflow for the duration of the grant, there have been several major challenges that have forced us to make certain changes in our approach to processing texts. The first was that, although we wanted to produce high-quality TEI documents, there were no best-practice models in the field of South Asian Studies for doing so. Indeed, the kinds of texts that we were working with had several features that were not directly addressed in the TEI guidelines, or admitted of several possible encoding solutions. The distinctive feature of our texts was the prominence of the commentary as a philosophical genre. We therefore had to find a solution for encoding a "base text," a "commentary," and possibly several "subcommentaries," as well as encoding the relationships between them, both at the level of structure, and at the level of references from the commentary to the base text. But even for more fundamental questions of text encoding, there were no standard practices, and no models that could be followed. How, for example, were we to represent a typical Sanskrit verse—metrically divisible into four "quarters," and often typographically set on two printed lines—in the vocabulary of the TEI? The fact that the TEI

guidelines allow for vastly different encoding practices can be gauged from the state of SARIT's text collection prior to the commencement of the grant: although all of the texts were, technically, valid TEI documents, they represent widely different approaches to text encoding, such that the "same" structures, such as verses or prose paragraphs, were represented by different kinds of markup throughout the corpus. A very practical requirement for proceeding with the text-encoding part of the project, then, was the formulation of a more restrictive set of instructions for encoding the specific kinds of texts that we were working with, in order to apply the TEI guidelines consistently.

The second major challenge was related to the first one, in the sense that a consistent approach to text-encoding was necessary not only for *producing* TEI documents, but also for *interacting* with them programmatically in a web-based interface. Our developers were trying to build an application that would display the texts that we produced, but in order to do that, they needed to know precisely how certain features of our texts would be encoded. So long as we did not settle upon a consistent approach to text encoding, and preferably produce a computer-readable instantiation of our approach (i.e., a more restrictive modification of the TEI P5 schema), the developers were building an application for a "moving target." Conversely, so long as the staff working on text encoding did not know the development trajectory of the web application, we were uncertain as to which aspects of the text would ultimately be "visible" to users in the web application. For example, we did not know whether the new version of the web application would support internal cross-references within a text, and therefore we were uncertain as to whether we should manually encode those cross-references in our TEI documents. As it turned out, the early versions of the web application required us to adopt some idiosyncratic encoding practices, such as assigning every single block-level element an xml:id. We revised these practices as development on the web application proceeded.

The lack of a consistent approach to text encoding posed a third and final challenge. We were sending texts to our double-keyboarding vendor in India to be typed in, but we were uncertain about precisely which features of the text would be represented in the final TEI document. This meant that we were uncertain about how to divide the encoding work between our double-keyboarding vendor in India and the text-processing staff in the United States. For example, if a printed edition uses bold-face text to mark quotations, then we might tell the vendor to enclose bold text in <q> … </q> tags (which we may or may not have replaced with other tags in subsequent text-processing). If, however, the printed edition uses bold-face text to mark both quotations and personal names, then we would have to decide whether to represent all the bold-face text in the same way, thus "flattening" the distinction between quotations and personal names, or whether to manually inspect every instance of bold-face text and decide whether it should be encoded as <q> … </q> or <name> … </name>. As it turned out, however, our double-keyboarding vendor proved to be quite unreliable when it came to consistently applying encoding throughout a text. Thus, although over the course of the project we asked our vendor to produce documents that were closer and closer to valid TEI, the necessity remained of checking every document manually and introducing more complex levels of markup. Our colleagues in Heidelberg switched to a different

double-keyboarding vendor midway through the project, Aurorachana in Auroville, Tamil Nadu, India. This vendor paid much closer attention to markup, and validated every document against a schema, although it took them longer to produce the texts. This resulted in a major difference between the Columbia and Heidelberg subprojects: the texts that they received from their vendor were closer to valid TEI, and required less manual postprocessing, whereas the texts that we received from our vendor were closer to raw text data, and required a great deal of manual postprocessing.

Under these circumstances, it might have made sense to begin the project by specifying precisely the kind of encoding we wanted to ultimately arrive at for every text. The problem, however, was that at the beginning of the project, the SARIT staff collectively had little experience of encoding Sanskrit texts in TEI. We could not credibly produce a set of encoding guidelines that ought to have been authoritative, not simply for our project, but for similar projects in the field, on the basis of our narrow experience, even if that represented the collective experience of producing digital texts of the field of Indology as a whole. In this sense, we had little choice but to proceed with encoding texts on the basis of our individual understanding of the TEI guidelines, and then to check our encoding decisions against each other, and revise our practice whenever we identified inconsistencies. One of our key principles at this stage, and indeed throughout the project, was to prefer "overencoding" to "underencoding." Given that we did not know which features of our texts would be supported in the web application, we preferred to create documents that had relatively high levels of markup, even if that markup would be "invisible" to the web application, instead of creating documents that had relatively simple markup just because we could count on the web application supporting it. Although this was a costly decision, in terms of time and labor, we continue to believe that it is important to separate the text data as such from the way that any given web application supports it: as we have learned, web applications come and go, but we continue to use the same text data. As a result of this decision, most of our texts are encoded with features that go far beyond what the SARIT web application, even in its most recent instantiation, is capable of interacting with. These features, which are detailed in the discussion of individual texts below, include detailed levels of structural hierarchy, systems of internal cross-references, and a critical apparatus. Our hope is that future applications, either developed in the context of the SARIT project or outside of it, will be able to leverage these features.

We proceeded in this fashion for the first year or so of the project, and then three of the project staff—McAllister, Olalde, and Ollett—met every couple of weeks to unify and document our encoding practices. The "SARIT Guidelines," which we consider to be one of the most important deliverables of the project, were publicly released in 2015. (See below, under "Promoting Standards," for the response of the South Asian Studies community to these guidelines.) There remain some major differences between the texts encoded before the guidelines were completed, and those that were encoded afterwards. For example, the Columbia subproject had not anticipated that line-breaks would be systematically encoded, and therefore we eliminated most line-breaks in the texts that were completed prior to 2015. For the most part, however, revising the texts that were

encoded previously was simply a matter of replacing certain tags with others, or "refactoring" certain parts of the document. After the release of the guidelines, the process of text encoding has posed more practical than organizational or conceptual challenges; it is simply a matter of introducing the markup specified in our guidelines into the texts supplied by our vendors.

This practical challenge, however, was far from trivial. As documented below, our subprojects each received tens of thousands of pages of typed-up Sanskrit texts. We needed to find a way to process this data into a valid and consistently-encoded TEI documents, with very limited human resources—usually just one quarter-time staff member. At the start of the project, we had planned to use a set of scripts to produce TEI data from raw text data. As it turned out, however, these scripts were not suitable for the text files that we received from our double-keyboarding vendors. One of the recurring issues was that TEI documents need to be encoded with absolute consistency, whereas printed editions—and the raw text files that are produced from them—do not have to be consistent in the same way. There are, for example, many cases of quotations that lack either an opening or a closing quotation mark, abbreviations that lack punctuation, labels that have parentheses in some cases but not in others. All of these are minor nuisances to the reader of a print edition, but they result in serious problems when trying to convert a text to an "ordered hierarchy of content objects" as TEI requires. The problems are compounded when a text has gone through multiple editions, or is issued in multiple volumes, as was often the case for the texts we processed in this project. In one of our texts, the notes were set in a completely inconsistent way, with footnotes that dated from an earlier edition of the text adopting one format (with Devanāgarī abbreviations) and footnotes that dated from a later edition adopting another (with Latin abbreviations). In the Columbia subproject, we were also constrained by a lack of expertise in automated text processing. None of the long-term project members had any background in computer programming. It was therefore a welcome change when one of our undergraduate assistants wrote a number of simple Python scripts to divided texts into structural units and to move footnotes from the bottom of the page into the main body of the text. Ollett subsequently used these scripts, with minor modifications, for all of the remaining texts. Much of the postprocessing, however, still had to be done manually.

Most of the texts that we chose to digitize were no longer protected by the copyright provisions that apply to scholarly editions. In a few cases, however, we sought and received the permission of the copyright holders to make the texts available online under a Creative Commons license.

The texts that the Columbia subproject produced for the project are as follows, in order of date of preparation. Note that all of these texts have been digitized, in the sense that raw text data has been supplied by our double-keyboarding vendors in India; what remains to be done, for the texts that are not yet completed, is the addition of TEI markup. Note that, unless specific encoding issues are noted below, all of these texts are encoded to the highest standards, with TEI markup representing the structural divisions of the text, its pagination, the notes of the editors and a critical apparatus, and often integrating corrections from the published corrections sheet (*viśuddhipattram*).

Kāvyādarśa (*Mirror of Literature*) of Daṇḍin, with the commentary *Ratnaśrī* by Ratnaśrījñāna. Completed in 2013. 287 pages. This is one of the foundational works of the tradition of poetics in Sanskrit, written around 700 CE, and SARIT presents it with its earliest surviving commentary (mid-10th century). In three chapters. Daṇḍin's text is in verse throughout; Ratnaśrījñāna's commentary is largely in prose, with several verse quotations.

Sarasvatīkaṇṭhābharaṇa (*Necklace of Sarasvatī*) of Bhoja, with the commentaries of Rāmasiṃha and Jagaddhara. Completed in 2013. 744 pages. One of Bhoja's encyclopedic works on poetics, composed in the first half of the 11th century, presented with two commentaries. In five chapters, in verse and prose.

Tantravārttika (*Commentary on the System*) by Kumārila Bhaṭṭa. Completed in 2016. 1036 pages. This text, an authoritative commentary on the Mīmāṃsā system of philosophy, covers some of the most important hermeneutical topics in that system. It was composed around the 7th century in a mixture of prose and verse, and is divided into chapters (*adhyāya*s), quarters (*pāda*s), and topics (*adhikaraṇa*s).

Daśarūpaka (*Ten Forms*) by Dhanañjaya, with the commentaries *Avaloka* by Dhanika and *Laghuṭīkā* by Bhaṭṭa Nṛsiṃha. Completed in 2016. 311 pages. This text is one of the most influential treatises on dramaturgy, written around the end of the tenth century. The *Daśarūpaka* itself is composed in verse, but Dhanika's near-contemporary *Avaloka* is in prose, and contains a large number of quotations, including quotations of theatrical texts, which made it a particularly complicated text to encode.

Nyāyasudhā (*Nectar of Principles*) by Someśvara. Completed in 2017. 1964 pages. This is a commentary on Kumārila's *Tantravārttika*, composed around the 11th century. Our source edition contains cross-references to page numbers of the edition of the *Tantravārttika* published in the Banaras Sanskrit Series.

Tautātitamatatilaka (*Adornment if Kumārila's Doctrine*) of Bhavadeva. Scheduled for completion in 2017. 870 pages. This is another commentary on Kumārila's *Tantravārttika*, composed around the 12th century. Unlike the *Nyāyasudhā*, this commentary clearly marks each topic (*adhikaraṇa*) of the *Tantravārttika*, and sets out Kumārila's own position against that of his philosophical adversaries, Prabhākara and his followers.

Śivārkamaṇidīpikā (*Jewel-Lamp to the Sun of Śiva*) by Appayya Dīkṣita. Scheduled for completion in 2017. 1094 pages. This is a subcommentary on Śrīkaṇṭha's commentary on the *Brahmasūtra*s, in which the South Indian polymath Appayya Dīkṣita (16th century) articulated his vision of a non-dualistic Shaiva philosophy.

Śṛṅgāraprakāśa (*Illumination of the Erotic*) by Bhoja. Scheduled for completion in 2017. 1630 pages. This is another encyclopedic work on poetics by the 11th-century king Bhoja. In contrast to his *Sarasvatīkaṇṭhābharaṇa*, however, the Śṛṅgāraprakāśa has a very complex structure, made up of lists within lists within lists. Our approach to encoding the text has been to represent these levels

of structure explicitly, which has proven to be a difficult undertaking; we may release the text with a "flatter" structure, at least until there is time to encode the hierarchical structure manually. Moreover, the text is in copyright, and we did not receive permission from the copyright holder to publish the digital text on SARIT until late 2016.

Nāṭyaśāstra (*Treatise on Theater*) of Bharata, with the commentary *Abhinavabhāratī* by Abhinavagupta. Scheduled for completion in 2017. 1678 pages. This is the foundational text of dramaturgy, composed around the middle of the first millennium CE, and Abhinavagupta's comprehensive commentary on it (ca. 1000 CE) is one of the most important expositions of poetics and aesthetics from premodern India. This text was to be the centerpiece of a collaborative editing project, discussed below, but the technical infrastructure of that project was never developed. The enormous inconsistencies of annotation of the source editions, moreover, necessitated manual corrections to nearly every single footnote, which has delayed the encoding of this text. The first of four published volumes has been completed according to our high standards for textual markup, but in order to release the following three volumes to the public in a timely manner, we may omit the text-critical markup for the *Nāṭyaśāstra* (but not for the *Abhinavabhāratī*).

Ślokavārttika (*Verse Commentary*) by Kumārila Bhaṭṭa, with the commentary *Tātparyaṭīkā* by Umveka. Scheduled for completion in 2017. 474 pages. This is the companion to Kumārila's *Tantravārttika*. It is composed in verse throughout, and organized into topics that deal primarily with the epistemological preliminaries of Mīmāṃsā's hermeneutical project.

Bṛhatī (*Long Commentary*) by Prabhākara, with the commentary *Ṛjuvimalā* by Śālikanātha Miśra. Scheduled for completion in 2017. 2258 pages. This is another authoritative commentary on the Mīmāṃsā system, composed in prose around the 7th century; Śālikanātha's subcommentary, the only that survives, was composed around the 9th century. Prabhākara and Śālikanātha represent a starkly different approach to Mīmāṃsā from that of Kumārila, and given that the structure of the *Bṛhatī* corresponds closely to the structure of the *Ślokavārttika* and *Tantravārttika*, we hope to provide a way of reading these texts side-by-side. As documented in our final semiannual report, however, a number of factors have delayed the production of TEI text of the *Bṛhatī* and the *Ṛjuvimalā.* These texts were published in four volumes, but the presentation of the text and commentary changes across these four volumes, and because of the fragmentary nature of the materials on which the edition was based—as well as, presumably, the tendency of the authors to skip from topic to topic—the published text does not cover all of the "topics" (*adhikaraṇa*s) of the Mīmāṃsā system, and it is often difficult to tell, just on the basis of a given printed page, to which book, chapter, and topic the text pertains. Hence, as in the case of the *Śṛṅgāraprakāśa* of Bhoja, we are considering publishing the text in a "flatter" format, at least until we are able to manually mark-up all of the relevant structural divisions.

We also have had double-keyboarded versions of the following Hindi texts prepared, which collectively represent a major portion of the corpus of classical Hindi literature, from the late sixteenth and seventeenth centuries: the collected works of the poets Keśavdās (*Keśav Granthāvalī*, 780 pages), Rahīm (*Rahīm Granthāvalī*, 80 pages), Sundar (*Sundar Granthāvalī*, 68 pages), and

Gaṅg (*Gaṅg Granthāvalī*, 185 pages). All of these texts should be converted to TEI format and appear in SARIT by the end of 2017. With the addition of these texts, SARIT will contain texts in three languages: Sanskrit, Prakrit, and Hindi.

In addition to these texts which we have digitized in their entirety, representing 13,451 pages of printed text, we have also digitized parts of two other texts, the anonymous *Kāvyakalpalatāviveka* (18 pages) and the second chapter of Hemacandra's *Kāvyānuśāsana* (*Teaching on Literature*), with his own commentary called *Viveka* (71 pages). We chose these texts because they contain close parallels to the *Abhinavabhāratī*, and we were expecting to be able to integrate them with the *Abhinavabhāratī* in a collaborative editing environment.

These figures do not include the texts that were prepared for SARIT by our colleagues in Heidelberg. The Heidelberg subproject has had one dedicated staff member, Liudmila Olalde, who has worked on text encoding over the past four years and produced dozens of high-quality TEI texts for the project; at first, the Heidelberg subproject focused on epistemological works by Buddhist authors, but have recently completed works by Jains and Brahmanical authors as well. They also do not include texts that were contributed by third parties to SARIT. While such contributions were once relatively rare, there are now a few important such texts in the corpus, and we expect to receive more contributions now that the SARIT Guidelines have been published.

These texts are stored in a public GitHub repository maintained by the SARIT organization. They are free to download, and available under a permissive Creative Commons license; we encourage users to use them, and modify them, for their own projects and research needs. We also hope that users will improve the texts, for example by correcting typos, or perhaps by adding new kinds of annotation. GitHub allows users to issue a "pull request" when they have made changes to files in a public repository: the maintainers of that repository can then review the changes and integrate them if they approve of them. So far, the only users who have used this feature have been members of the SARIT organization, but in principle, everyone is welcome to improve the texts in this way. Occasionally the SARIT staff receive lists of corrigenda to SARIT files from interested scholars, which we then integrate into the texts. The Git repository also forms the basis for the SARIT data collection, which is rebuilt and uploaded into the SARIT web application whenever the Git repository is updated; see below for more on the web application. There are many use-scenarios for the TEI documents provided by the SARIT project. Users can use XSLT or XQuery to parse the document and query it like a database: you can, for example, retrieve all of the quotations of a certain work by a certain author, and order them by occurrence or alphabetically; you can easily generate indices of verses, or lists of authors or works mentioned, provided that these have been marked up in the TEI document; you can isolate the commentary from the base text or *vice versa*; and, of course, you can search the document for keywords.

B. Development of a Web-based Interface

Most users in the scholarly community do not prefer to interact directly with the TEI documents that SARIT makes available on its public GitHub repository. They prefer to interact with the documents

in a more “human-readable” format. For reading the documents, this might be an EPUB or PDF format. For searching the documents, it is likely to be a web-based search application. For these reasons, a web-based application for interacting with TEI documents has always been part of the vision of SARIT. Before the start of the NEH-DFG grant, SARIT’s website used an implementation of Philologic, an open-source web application developed by the University of Chicago that had two crucial features: it allowed users to see an HTML rendering of the source TEI document, and it provided a search function. Philologic, however, had several major limitations. It was not in active development at the time, and its code base did not allow the SARIT staff to easily make modifications to the application. It also used CSS to render the TEI elements directly, rather than transforming them into HTML first with XSLT or XQuery. This imposed some major restrictions on the rendition of SARIT documents in the web application, and made it impossible to use a lot of “Web 2.0” features, such as dynamic modals or responsive web design. What’s more, additions to the text corpus caused SARIT’s deployment of Philologic to fail in early 2014, which made its replacement all the more urgent.

The responsibility for developing the web application primarily fell to the Heidelberg subproject for institutional reasons. Columbia offers little support for software development. There is an in-house software development office, called CDRS, but they had very limited resources and no substantial experience with the kinds of text-encoding projects that SARIT represented. By contrast, the University of Heidelberg has an in-house development office, called the Heidelberg Research Architecture (HRA), that had close ties with eXist Solutions, the firm that develops eXistDB, which is one of the more widely-used database programs in the scholarly world and particularly within the world of digital texts. Staff at the HRA were willing to work on a rewrite of the SARIT web application, using eXistDB as the underlying architecture. The SARIT team approved of this approach, given the fact that eXistDB is open source, powerful, and relatively easy to install and maintain. Some of these developers, moreover, worked closely with Wolfgang Meier, the leader developer of eXistDB, who came to be involved with the SARIT project from an early date. Thus the developers who worked on SARIT not only used eXistDB, but were actively involved in maintaining and updating it.

This close relationship with the developers of eXistDB paid off relatively quickly. At the beginning of the project, we realized that the index used for searching the texts, both for the earlier Philologic setup and for the proposed eXistDB setup, presupposed the division of every text into words. This proved to be a major problem for Sanskrit texts, where words are not systematically separated by spaces. The lack of systematic word-division meant that users could not easily search for words, because the “real” words were indexed as substrings of apparent “words” which were in fact simply groups of characters separated by a space. Wolfgang Meier proposed indexing the SARIT texts, instead, using NGrams: rather than indexing groups of characters separated by a space, the database would index adjacent groups of three characters. Meier programmed the NGram index and included it in the public release of eXistDB, and we used this index in both the development and public deployment of the SARIT application from 2014 until 2017. The use of an NGram index

is one example of how the challenges posed by South Asian textuality—in this case, the use of a syllabic script and the unfamiliar practices of word-division that this implies—lead to major changes in the way that people think about and interact with digital texts. Rather than treating them as collections of words, it became necessary to treat them as collections of characters.

The SARIT website was launched as an eXistDB application in May 2015 after about a year of development and testing. The new website provided the possibility of searching all of the SARIT texts, as well as those by specified authors, using either the word-based index or the new NGram index. The other important feature of the new website is that Jens Petersen, formerly of the HRA, wrote a set of XQuery routines to transform the TEI documents to HTML output. These routines were more powerful than the CSS-based rendition of Philologic, and could be customized.

Several major challenges remained even after the public launch of the SARIT website. One was related to the fact that the SARIT corpus contained documents in both the Devanāgarī script and in Latin transliteration (using the International Alphabet for Sanskrit Transliteration or IAST). This "two-script" model was a legacy of the way the SARIT corpus was built up, first by scholars typing up their own texts in IAST, and subsequently by projects having texts typed up in Devanāgarī by double-keyboarding vendors. It was, however, important to the project staff to make SARIT as script-neutral as possible. In other words, we did not want to constrain users to browse texts, or search texts, in either IAST or Devanāgarī. Indeed, since Sanskrit has traditionally been written in about a dozen regional scripts, we thoughts that SARIT should ultimately support the searching and display of texts in any Indic script of the user's choice. Ollett and Petersen modified SARIT's search interface to accept queries in either script, but this modification proved untenable: because of the different nature of the Latin and Devanāgarī scripts, the former being alphabetic and the latter being syllabic, the logic of searching them was quite different, and it was not possible to simply convert search strings, especially search strings with wildcards, into one or the other script with the same behavior; moreover, some characters which are single codepoints in the Devanāgarī script are represented by two codepoints in IAST (such as "kh" and "ai"), which generated additional issues for searching texts in both scripts using the same search query. In 2016, Claudius Teodorescu, a developer at the HRA, had the idea of indexing every text in the SLP1 transliteration scheme, developed by Peter Scharf and Malcolm Hyman, which provided a one-to-one mapping between Devanāgarī and Latin characters. This meant that users would be able to search in either IAST or Devanāgarī and retrieve results from texts that were prepared in either script, bringing SARIT much closer to the ideal of script neutrality. Teodorescu introduced the "single index" version of the SARIT application in 2016, and it turned out to solve a host of issues related to unexpected behavior of search results when matching IAST to Devanāgarī or *vice versa*. The "single index" is indeed an elegant solution to a problem that is endemic to South Asian Studies and has only grown more severe as Unicode has found greater acceptance: the use of multiple scripts to represent what is underlyingly the "same" text. It also represents a key value-added feature of the SARIT website. For a long time, searching documents through a web interface did not offer any particular advantages over searching documents on one's computer with a command-line program such as grep. Now,

however, users of the web interface can search documents in both scripts at once. In the final phases of the project, we are implementing a transliteration-on-the-fly feature, which will allow users to read texts on the SARIT website in a script of their choice. This feature, like the transcoding routines that result in the SLP1 index, is based on a Java library provided by the Sanskrit Library, directed by Peter Scharf. It represents another significant step in the direction of script neutrality.

An additional advantage of Teodorescu's work on the SLP1 index was that, between 2013 and 2015, Apache Lucene—which eXistDB uses for its full-text index—began to support indexing using NGrams rather than word tokens. Thus Teodorescu's version of the application uses the NGram index built into Lucene, rather than the separate NGram index that Wolfgang Meier had designed specifically for eXistDB. This has streamlined the application somewhat, and also improved its long-term sustainability by using Lucene, which is well-documented and well-maintained, for all of the core indexing and search functions.

From the beginning of the grant period, the Columbia subproject insisted that the SARIT web application must have collaborative editing functionality. This means that the web application, instead of simply indexing and displaying TEI documents that were loaded into it from another source, would actually be able to modify the TEI documents themselves. Users could therefore propose corrections to texts *within* the application, rather than by issuing a pull request on SARIT's public Git repository. In principle, this would have allowed a much more interactive approach to text encoding: instead of relying on a small staff of people trained in TEI to prepare every part of every document, we could rely on a wider network of people who would be allowed to make interventions in the documents in an intuitive web-based interface. The use-case that we envisioned for this functionality was a collaborative reedition of the *Abhinavabhāratī*, one of the texts that the Columbia subproject was preparing for SARIT. The *Abhinavabhāratī* is precisely the kind of text for which collaborative editing makes a great deal of sense: it is large, and spans subjects on which no one scholar has a complete grasp today, and the manuscripts are so lacunose and corrupt that it is possible for intelligent readers to make *ex ope ingenii* conjectures on nearly every single page. Pollock and Ollett prepared a proposal for this subproject, tentatively titled *Abhinavabhāratī Online*. It involved adding an authentication system to SARIT, so that users with certain permissions could log in and see manuscript images of the *Abhinavabhāratī* (collected by Pollock from the University of Kerala Research and Manuscript Library, Trivandrum, India, with his own resources) alongside the TEI text, which is based on the critical edition published by the University of Baroda. (Pollock had already received permission from the University of Baroda and the University of Kerala to proceed with this subproject.) Authenticated users could then make a number of annotations to the TEI text, broadly classified into "semantic" annotations (representing the identification of quotations, personal names and titles of works, and so on) and "text-critical" annotations (representing variant readings, conjectures, or textual notes). We discussed this proposal in detail, both with the staff of the Heidelberg subproject and the developers at the HRA. At the time, and still now, there was considerable general interest in a "TEI editor" that would allow users to edit TEI documents on the web through an intuitive, schema-aware interface. There was also considerable

interest in annotation, which has only blossomed since 2014 with the widespread adoption of "linked data" formats. The technical challenges of implementing such an editor, however, were considerable. Given the requirement to preserve the integrity of the text data, and to maintain the overall Git-to-eXistDB workflow of the SARIT project, the developers suggested using standoff markup, where annotations would be stored separately from the main text and indexed to particular positions in the text stream. At the time, however, standoff markup for TEI documents had not been successfully implemented in any project known to us, and there were considerable practical difficulties in combining a schema-aware TEI editor with standoff markup and synchronization with GitHub. The departure of Jens Petersen from the HRA, and Wolfgang Meier's other commitments, made further development on these features unfeasible. The Columbia subproject actively sought to identify existing tools, but none was adequate; it looked for other developers to take up the task, but found none who were available during the remainder of the NEH-DFG funded project. To our disappointment, work on the collaborative editor could not proceed further.

In the meantime, however, Wolfgang Meier's other projects led him to reconceive the way that the SARIT application should work. A project focused on early modern printed texts in English, based at Northwestern University, led Wolfgang Meier to work on a new processing model for displaying TEI documents in a web browser, originally based on the "TEI Simple" subset of the TEI P5 guidelines. The result of this work was an application called "TEI Publisher," which allowed users to create their own applications using a given data collection (i.e., a set of TEI documents) and an ODD file that specified the rendition of the TEI elements. The use of an ODD ("One Document Does it all") is a radical departure from the earlier way of rendering TEI elements in the browser, which was to write XQuery to transform the TEI into HTML. With the new processing model, one simply specifies how the TEI elements should behave directly in the ODD file, which also acts as a schema. The ODD file is then used to automatically generate intermediate files for producing output in a variety of formats, including HTML, PDF, and EPUB. At the insistence of the SARIT staff, the processing model uses LaTeX in addition to FO to produce PDF documents. Currently the SARIT web application only partially implements the new processing model: Wolfgang Meier rewrote the application code for SARIT before completing the TEI Publisher application, and hence the current version of SARIT still uses hard-coded XQuery to render TEI elements into HTML. The application, however, is scheduled to be rewritten once again to make use of the new TEI Publisher libraries, and the new version of the application will allow the SARIT staff to alter the appearance of texts in the corpus by modifying the ODD. As of the time of writing (July 2017), the SARIT project has published an ODD file that contains both the schema and the documentation for the SARIT texts in our public GitHub repository, although the development branch of the web application does not yet use this ODD file for displaying the SARIT texts.

In the last months of the NEH-DFG funded project, we have also taken steps to ensure that the SARIT project members can maintain the web application themselves, i.e., without the assistance of the developers who have, during the four years of the project, been responsible for tasks such as configuring the server, updating the data repositories, reindexing the database, and so

on. Several project members, including Ollett, run the SARIT web application on their local machines, and have experience with installing and running eXist-DB applications on a web server. We are hopeful that the SARIT project staff will be able to continue not only running and updating the SARIT web application, but also developing it further, and in particular, making changes to the display of texts—either in HTML or in PDF format—by specifying behaviors in the ODD. The development team has been actively working on the rewrite of the application code (in the SARIT-PM, or "processing model," branch), and the project members have been waiting for these tasks to be completed before further refining the display of texts in the application.

Thus the web application has gone through several different incarnations. Although development has not proceeded in the way that the members of the Columbia subproject would have preferred, and in particular we are disappointed that the development of a collaborative editor was not feasible at this time, we are satisfied that the project's resources have gone towards solving fundamental problems with searching and browsing Sanskrit texts, prepared in two different scripts, in a web application. We are also confident that these solutions will both contribute to the longer-term sustainability of SARIT and find application beyond the SARIT project. The capacity to search texts in both scripts, and the transcoding architecture underlying this capacity, represents an important new tool for researchers. And the transition to an ODD-based processing model, where the human-readable output results from a single set of behavior specifications rather than a collection of difficult-to-maintain XQuery code, should make it easy to adapt the SARIT web application to new kinds of texts, on the one hand, and indeed encourage users to define their own preferred behaviors.

C. Bio-bibliographic Database

One of Pollock's previous projects, Sanskrit Knowledge Systems on the Eve of Colonialism (SKSEC), involved the production of a bio-bibliographical database that focused on early modern authors of Sanskrit, the texts that they wrote, manuscripts of these texts, and secondary scholarship. The database was in MySQL format. When Pollock moved from the University of Chicago to Columbia University, the database had to be moved, as a result of Columbia's inability to support it. University College, London, was chosen, since a collaborator on the SKSEC project was appointed there, but by 2013, with the death of the programmer at UCL, there was nobody left to maintain the database, and it was basically defunct. The NEH-DFG grant provided an opportunity to revive the SKSEC database in the context of the SARIT project. We hoped, in particular, to be able to fix certain corruptions in the database and to integrate the bio-bibliographical data from SKSEC with the text data from SARIT.

In 2013, Ollett received the SKSEC database files from technicians at UCL and began to correct some of the systematic errors that were introduced when the database was converted to a Unicode encoding several years previously. Around the same time, however, we were contacted by Yigal Bronner, at the Hebrew University, Jerusalem, who was planning to create a bio-bibliographical database centered on the South Indian intellectual Appayya Dīkṣita (1520–1593 CE).

Because the database Bronner envisioned would have overlapped to some extent with the SKSEC database—to which Bronner had previously contributed—we began to discuss the possibility of merging the two projects, and in particular, of importing the relational data of the SKSEC database into the Drupal-based platform that Bronner had designed with Amir Siman Tov, a software developer based in Israel. Over the next year, Ollett helped Bronner and Siman Tov create scripts to import the SKSEC data into the new database. The *Pandit Project*, as it is now called (for "Prosopographical Database of Indic Texts"), now hosts all of the erstwhile SKSEC data, consisting of 8820 rows of data in 20 database tables.

The *Pandit Project* has grown with successive imports of data from other sources, as well as from new contributions from users since the public release of the project in the summer of 2016. As of May 9, 2017, it contains 50,714 total records, covering 9,442 primary texts in Sanskrit and other South Asian languages, 3,869 persons, 2,130 manuscripts, 113 places, and 35,160 works of secondary scholarship. There is now a small community of users who regularly contribute and revise data, and the depth and breadth of coverage will continue to improve, given that the *Pandit Project* now functions as a data repository and publication venue for several funded research projects.

D. Promoting Standards and Practices for Digital Philology

What has distinguished the SARIT project from its beginning from other projects concerned with digital texts in South Asian Studies is its concern with standards. By "standards" we mean approaches to producing, maintaining, and distributing digital texts that other projects can follow, and, through their adherence to these standards, produce data which is broadly consistent with the data the SARIT project has produced and which will work with some of the same tools, such as the SARIT web application. Generally, when digital texts projects speak of "standards," they refer to the Text Encoding Initiative and its guidelines (http://www.tei-c.org/Guidelines/P5/), which have indeed set the standard for scholarly digital texts across a range of disciplines. In the context of the SARIT project, however, "standards" means something both broader and narrower than following the TEI guidelines, as the project members have come to learn over the past several years in their interactions with other projects.

The broader vision of standards implies a commitment to producing *open text data.* By this we mean data that is, in the first instance, accessible and downloadable as text documents, and which is "open" in the sense of available to anyone with an internet connection, without specialized technical knowledge, and under permissive licenses. A corollary of this commitment is that the text data comes first, and databases second: we want to avoid the persistent problem in digital humanities of data being "stranded" in databases that are no longer actively maintained. Hence the SARIT project has always kept its data repository separate from the development of the web application, which is an XML database, and we have designed the web application in such a way that the database can never make changes to the text data which are not immediately reflected in the data repository. The SARIT project is by far not the first digital text project to use GitHub to store

and maintain its text data, but this arrangement does reflect our commitment to open text data. We also believe that text data which is "locked up" in a database is not sufficiently open, even when it is made available under permissive licenses, since members of the public are rarely able to retrieve the data under such circumstances, usually for lack of database privileges. We have found, anecdotally and through surveys, that one of the things that the community likes about SARIT is that they can easily download the text data and use it however they wish.

The narrower vision of standards means that *TEI is not enough*. We learned very quickly that the TEI guidelines often present many different encoding strategies for the same problem. This tendency is one of the TEI's strengths, because different encoding strategies are appropriate for different projects. But it was an obstacle for us, in the beginning, because we had a variety of texts that were prepared by different people who had made different encoding decisions. All of the texts were "TEI-conformant," in the sense that they validated against the schema supplied by the TEI. And although some texts clearly violated the spirit of the TEI guidelines, most of the texts represented a genuine attempt to adapt the TEI guidelines to the practical needs of Sanskrit texts. Those who contributed texts, included members of the SARIT project themselves, had decided to deal with certain encoding issues in different ways. These divergent encoding practices presented a problem in 2013, when we began to build a web application to search and display the SARIT texts. If we were going to have a consistent approach to interacting with texts, we needed the texts themselves to be relatively consistent. Thus, in the absence of a set of guidelines in common use for applying the principles of TEI to Sanskrit texts, we decided that the SARIT project itself must produce such guidelines. Relatedly, we decided that the SARIT project should also publish a "restricted schema" based on the TEI schema, although for most of the duration of the NEH-DFG funded project, there was a greater need for human-readable guidelines than for machine-readable schemas.

Three of the core project members—McAllister, Olalde, and Ollett—wrote a set of guidelines for applying TEI markup to Sanskrit texts, which has since been published on the SARIT website as the "Simple Guidelines." These guidelines were called "Simple" because they dealt with only the most basic encoding scenarios. In other words, if anyone were to attempt to represent a Sanskrit text as a TEI document, he or she would almost certainly encounter situations for which our documentation provides detailed guidance. These situations include the encoding of metrical units within verse texts, the representation of a "base text" and a commentary, and notes, as well as general guidance for producing a valid TEI document. The "Simple Guidelines" anticipated a set of "Advanced Guidelines" which was to deal with less common but more complicated encoding scenarios, such as the representation of variant readings in a critical apparatus and other types of annotation. The "Advanced Guidelines" were drafted in anticipation of support, within the SARIT web application, for these "advanced" features, and in particular, in the expectation that users would be able to make interventions into the text through a web-based editor that would be stored in the SARIT texts themselves as text-critical annotations in a TEI-compliant format. Those features were not built into the application in the present funding period, and hence the "Advanced Guidelines"

were never published: we decided that it did not make sense for SARIT to publish standards for levels of encoding for which the SARIT application did not offer robust support, especially since there was, and remains, a great deal of debate among the TEI community about how annotations ought to be stored. (See the discussion of stand-off markup above for further details.)

One way to disseminate the standards and practices favored in the SARIT project is to lead training workshops for scholars—including students—in the field of South Asian Studies. Our colleagues in Heidelberg have, for example, offered a seminar at the University of Heidelberg devoted to digital research methods in Indology, which introduced participants to the fundamentals of TEI and the most common ways of interacting with text data. The Columbia subproject did not undertake such activities, apart from training the personnel who worked with us, although it makes good sense to organize such events in the future.

At this moment, it seems that SARIT has been successful in its advocacy of the TEI as a standard for digital versions of Sanskrit texts. Among our colleagues in the field of South Asian Studies, the "SARIT approach" to producing digital texts has caught on. We know of at least three major projects that are committed to producing text data in accordance with SARIT's interpretation of the TEI guidelines, and which have explicitly referred to SARIT in their funding applications. The earliest to do so was a project led by James Mallinson, at the School of Oriental and African Studies in London, called "The Haṭha Yoga Project: Mapping Indian and Transnational Traditions of Physical Yoga through Philology and Ethnography," which has received funding from the European Research Council.[5] One of the primary goals of that project is to produce digital critical editions of Sanskrit texts that are significant for the history of Haṭhayoga, and they are using the SARIT guidelines in the preparation of those editions. Another project which explicitly follows the SARIT guidelines is "Documents on the History of Religion and Law of Pre-Modern Nepal," led by Christoph Zotter of the University of Heidelberg, and funded by the Heidelberg Academy of Sciences.[6] This project has already produced thousands of TEI editions of legal and administrative documents that are kept in the National Archives in Kathmandu. The use of the SARIT guidelines in this context is especially encouraging, given that the vast majority of the documents with which the project is concerned are in Newari rather than Sanskrit, and hence we can be confident that the SARIT guidelines can be adapted, or extended, to cover texts in other Indian languages besides Sanskrit. Finally, we can mention "Age of Vedānta" project, directed by Ajay Rao of the University of Toronto and Lawrence McCrea of Cornell University, which has so far received funding from Canada's Social Sciences and Humanities Research Council. This project is dedicated to uncovering the history of Vedānta, or non-dualistic thought, which dominated philosophical and religious discourse in early modern India. The project will provide for the digitization of important Vedānta text and contribute them to SARIT. This brief survey does not include projects that the members of SARIT have undertaken themselves, either with or without funding, such as Ollett's corpora of Sātavāhana Inscriptions and Prakrit poetry.[7]

5 http://cordis.europa.eu/project/rcn/199664_en.html
6 http://www.haw.uni-heidelberg.de/forschung/forschungsstellen/nepal/index.de.html
7 See, e.g., http://37.252.124.228:8080/exist/apps/SAI/browse.html.

One final example of SARIT's success in developing and adhering to standards in the production of Sanskrit texts happened, serendipitously, just as the NEH-DFG project was drawing to a conclusion. As noted above, SARIT used the Philologic software for its web application until early 2014. At that time, it was not clear that Philologic would be developed further, and we switched to eXist-DB. Recently, however, the newest version of Philologic was released by the ARTFL Project at the University of Chicago, and at the suggestion of our colleague Tyler Williams, the ARTFL Project pulled the TEI data for SARIT from our GitHub repository and loaded it into Philologic. As it turns out, the "out-of-the-box" installation of the SARIT texts worked: Philologic, which was not developed with Sanskrit texts in mind, expected to receive data in TEI format, and the SARIT data conformed to its expectation.[8] This is a good sign for integration between digital texts projects in the future: if we do indeed adhere to standards in our individual projects, it will become easier to share data, and to avail ourselves of the tools that others have developed.

3. Looking Toward the Future

At the time of writing, the major goals that the SARIT project set out to accomplish in its funding application have been realized: SARIT has evolved from a relatively small collection of texts into a sizeable archive, representing about a hundred thousand pages of text data; it has added, besides epochal texts such as the *Mahābhārata* and classics such as the *Arthaśāstra* and the *Laws of Manu*, the foundational texts of at least three major traditions of Indian thought—Buddhist logic and epistemology (*Pramāṇavidyā*), Vedic hermeneutics (*Mīmāṃsā*), and poetic theory (*alaṃkāraśāstra*); its encoding has evolved from a "by hook or by crook" approach to TEI to a consistent, principled, and well-documented interpretation of the TEI guidelines; the SARIT web platform has evolved from a fragile Philologic database to a robust eXistDB application, and in the process, has added useful features—like two-script searching—that are available nowhere else. And even in the field of digital humanities, when four years is more than enough time for technologies to become outmoded, SARIT remains at the forefront of innovation, as attested by its soon-to-be-implemented "processing model" and the use of ODD documents for schemas, documentation, and presentation. The standards and practices that the SARIT project has developed for itself have been relatively widely adopted in the field of South Asian Studies. Even apart from the handful of projects that explicitly refer to SARIT and base their contributions on SARIT's guidelines, we believe that SARIT can take some credit for the increased awareness of TEI, and the growing sense that it ought to be a core data format for digital philology, in South Asian Studies.

Since none of the core project members had been involved in a funded digital humanities project in the past, it is no surprise that we learned a number of important lessons in almost every domain, in general areas such as project management, and in specific technical areas such as the viability, or lack thereof, of stand-off markup. In the following, we will review some of these lessons, which we believe will be important starting points for future digital humanities projects in the field of South Asian Studies, including for future rounds of funding and development for SARIT.

8 See http://condorcet.uchicago.edu/philologic/sarit/.

It is of utmost importance to establish a relationship with professional software developers, to communicate often and clearly with them about the goals of the project, and to consistently review and discuss their progress towards those goals. We benefitted enormously from the University of Heidelberg's arrangement with eXist Solutions, which allowed us to work with the very people who wrote eXist-DB, who were in turn very responsive to—and interested in—the particular needs of the SARIT project. We might, however, have wanted to have a clearer understanding, at various stages of the project, of which version of the application represented the most recent changes to the code base. It was sometimes the case that we tested a feature that had been superceded in a more recent rewrite of the code, or conversely, we weren't aware of a new feature that had been added. We used the open-source project management tool Redmine, but outside of the SARIT project, several of us have found GitHub's "issues" function to be useful, in that it integrates discussion of issues with changes to the code base (represented as Git commits). Moving more of the development to GitHub would also have the advantage of allowing project members to track the changes in the code base more easily using Git's system of branches and commits.

We initially underestimated how important it would be to transform the text data programmatically in order to achieve the desired TEI markup. None of the project members had a formal background in computer programming. About halfway into the project, the Columbia subproject hired an undergraduate assistant, Sireesh Gururaja, who wrote a number of Python scripts to accomplish basic tasks, like moving footnotes from the bottom of the page into the location of the footnote marker in the text, which Ollett had earlier done using a combination of regular expression line editing tools and manual replacement. This made us aware of the necessity —not merely expediency—of involving collaborators from a computer science background, in addition to, or even in preference to, collaborators from an Indological background. The tension between technical and subject-area expertise is, of course, a common theme in the digital humanities. We had assumed, at the beginning of the project, that knowledge of Sanskrit was a necessary precondition to contributing in a meaningful way; at the end of the project, it is clear that this is not the case. The practical challenge we experienced, over the course of the project, was finding people with technical expertise that were willing and able to work on SARIT. The tasks were "too easy" for computer scientists and "too hard" for Sanskritists. The Heidelberg subproject had relatively more success in finding technicians, developers, and assistants with the requisite technical knowledge, and in hindsight we can account for the difference in two ways. First, the University of Heidelberg provided dedicated support to the project through its digital humanities division, the Heidelberg Research Architecture, whereas Columbia University provided no support whatsoever through its own digital humanities initiatives (an asymmetry of which we were fully aware when we submitted the applications to the DFG and NEH). Second, the Heidelberg subproject organized a seminar on digital research techniques, which increased the visibility and repute of the project in Heidelberg, whereas the Columbia subproject remained relatively unengaged from, and thus relatively unknown to, the community of digital humanists at Columbia.

As noted above, one of the challenges of a project like SARIT is proposing tasks that are difficult enough to be interesting to potential collaborators with a digital humanities or computer science background, and at the same time, not too difficult to realized within the course of a funded project. In the past four years, SARIT has, with very good reason, gone after low-hanging fruit. Goals that were deemed to be too ambitious, especially from a technical perspective, were postponed until the core functionality of SARIT—the searching and display of Sanskrit texts—was completed. We are now, however, in a position to argue that what is good for SARIT is good for the digital humanities community, or the natural language processing community, or the optical text recognition community, and so on. That is to say, we have now built up a corpus of Sanskrit texts that are encoded with a high degree of consistency and tagged in such a way that they can easily be linked to page images of the printed edition. We ought to encourage scholars to use this data in whatever way they can imagine, and we ought to team up with scholars in other disciplines, such as computer science, and suggest certain ways of using the data. Ideas that have occurred to us in the past are: the perennial problem of word-analysis in Sanskrit, which might be solved by training machine-learning algorithms on the SARIT corpus; the persistent problem of Devanāgarī OCR, which might be solved, once again, by training state-of-the-art OCR algorithms on the page images of the SARIT corpus and checked against the double-keyboarded texts; the possibility of applying techniques of named entity recognition, or topic recognition, to the SARIT corpus. These are all techniques that have been applied profitably in other fields, and SARIT can help South Asian Studies to "catch up." Of course, these techniques would also greatly increase the usefulness of the SARIT corpus and allow for its rapid expansion in the future.

We believe that three features, in particular, are well within SARIT's reach. While we did not have the resources to implement these features during the NEH-DFG funded project, it seems reasonable, from our current perspective, to believe that they can be implemented in the next round of development. They are collaborative annotation, a canonical reference system, and text-image alignment. We single out these three features not only because they would make SARIT a unique and indispensible resource for scholars of South Asian Studies but also because they represent SARIT's relationship to emerging standards in the digital humanities as a whole.

Collaborative annotation has been on the agenda since the beginning of the NEH-DFG project. It would make SARIT useful not only for individual research, but for collaborative projects and for teaching. We now recognize that our initial insistence on "collaborative editing" was slightly too narrow, and that the more common use cases for collaborative annotation don't involve interventions *into* the text, but comments *about* the text. Both types of annotation are relatively difficult to represent and store within a single TEI document, but the development of "Linked Data" standards in the past several years has meant that it is now possible to represent annotations in a consistent and persistent fashion, without "clogging up" the target of those annotations. When we discussed collaborative annotation with our developers in 2014, we agreed that SARIT could be pathbreaking digital humanities project if this feature were successfully implemented; since then, a number of other projects have made progress in this area, but using traditional TEI methods rather

than Linked Data methods.[9] One of the benefits that collaborative annotation would have is bridging the gap between the "power user" of SARIT who interacts directly with the TEI documents—and the number of such users is perhaps half a dozen—and the "ordinary working Sanskritist" who is not comfortable with TEI or XML, but nevertheless has important contributions to make to the texts in the SARIT archive.

In hindsight, text-image alignment seems like one of the "low-hanging fruit" that the SARIT project might easily have accomplished during its four years of NEH-DFG funding. After all, our workflow has started with page images of a given edition of a text and ended with a TEI document that represents a faithful transcription of the text of the edition and includes information about page- and line-breaks of that edition. To align the text and images, we would simply have to store the page images on the server, and have the page break element in the text "point" to the relevant page image. SARIT's texts, however, were supposed to be encoded to such a high standard that consulting the printed edition would be unnecessary, except in unusual circumstances, and thus we did not plan to provide the page images of the edition alongside the text. But no digital text is perfect, and even texts that are 99.9% accurate will contain one mistake every thousand characters. In order to make it easy for SARIT's users to check the digital text against the edition, we ought to provide page images; in order to make it easy for them to submit corrections to the text, we ought to provide annotation functions. And as noted above, a corpus of aligned text and images will be very useful to train OCR engines. Once we provide support for viewing page images alongside the digital text, as many other digital texts projects do, it should be relatively easy to link a text with images of manuscript. This would mean that SARIT could be used, along with other tools, for the "front line" philological work of manuscript transcription and collation. In the area of text-image alignment, too, technologies and standards have advanced considerably since the NEH-DFG funded project began. The IIIF standard has been adopted in a wide range of projects, bringing with it a range of tools (such as the Mirador image viewer) and services (such as the Loris image server), and it makes good sense to integrate SARIT with this standard.

Finally, there is the outstanding issue of a reference system. In the field of South Asian Studies, there is no "canonical" reference system such as that used by the TLG for Greek literature. Scholars still refer to the structural divisions, and sometimes also the page- and line-numbers, of specific editions: a reference such as *Mahābhārata 1.4.15.2* can only be "decoded" if one knows the edition to which it refers. There are, moreover, multiple conventions in place for referring to works; one never knows, without looking it up in a list of abbreviations, whether an abbreviation such as *MaBhā* refers to the *Mahābhāṣya* or the *Mahābhārata*. The SARIT project is in a unique position regarding the issue of a reference system. First, the texts are encoded in such a way that the structural and typographic divisions of the source editions—books, chapters, verses, as well as pages and lines—are represented in the TEI documents, and these divisions are of course interpretable as belonging to a particular edition or "instantiation" of the text through the TEI metadata. These divisions are already presented in the SARIT web application (which offers both a

9 The most promising project of which we are aware is the CWRC-Writer; see http://www.cwrc.ca/projects/infrastructure-projects/technical-projects/cwrc-writer/.

page-by-page view, and a view based on structural divisions). Second, the SARIT corpus now contains texts with references to other texts which have been encoded in a systematic way. Although currently we provide no way of "decoding" such references, it is obviously desirable to do so: thus a reader who sees a reference to one text within another might want to follow the reference to see, for example, the context of the citation. Finally, the SARIT corpus now contains a large enough number of texts, especially within the areas of its specialization during the NEH-DFG project, that, for a given textual reference, it is quite likely that the text referred to is either present in SARIT or will be present in SARIT in the future. The texts are not currently identified by a Uniform Resource Identifier (URI), since there has not been a demand for such identifiers in the past, but with a few small changes they can be so identified. As a result, SARIT is particularly well-poised to act as a "name authority" for Sanskrit texts: we can, in other words, publish a list of of SARIT texts, the abbreviations which ought to be used for those texts, and the reference scheme or schemes that are used in the SARIT texts to locate particular passages, and on the basis of this information, any arbitrary reference to a text present in SARIT, e.g. *MaBhā.1.4.15.2*, can be automatically "decoded" and the corresponding text can be displayed in the browser (or served through an API). Nor is there any reason to stop at the texts that are present in the SARIT corpus: the authority files maintained by SARIT could function as a general reference system for Sanskrit literature. Sanskrit scholars might be skeptical of the need for such a reference system—after all, the field has survived thus far without one—but digital humanists will immediately understand why such a system is necessary. Reference systems are needed to create links between sets of data, and hence they can also facilitate the sharing of data between projects. In fact, two of the projects that are closely affiliated with SARIT—the EAST bio-bibliographical database, and the PANDIT prosopographical database—will need ways of referring not only to *texts* that are available on SARIT, but particular passages of texts that are available on SARIT. Our model in conceiving such a reference system has been the CTS (Canonical Texts Services), a standard developed by Classicists, and once again, it should not be too difficult to implement a similar system for SARIT, especially since the CTS has been implemented in eXist-DB.

Finally, as Dominik Wujastyk's presentation at the SARIT conference in Vienna made clear, what users want from SARIT is not a single web application that "does it all," but a useful tool that they can use in conjunction with other tools. It is essential, in other words, to continue to develop SARIT with an appreciation of its unique role in an "ecosystem" of other resources and projects in South Asian Studies. Thus far, one of SARIT's key roles has been to promote the uptake of standards in a field where it is not immediately obvious why such standards should be followed, and we believe it would be appropriate, and in keeping with SARIT's mission, to demonstrate the full range of possibilities that the adherence to those standards opens up. Thus, while we will continue to promote the use of TEI encoding, we plan to offer more detailed guidance about such subjects as text-critical annotation, cross-referencing, and the alignment of texts (e.g., when there are multiple commentaries available on the same passage of a given text). We will also promote the use of consistent naming and referencing systems, as outlined above. Finally, we will promote the ideal of *open text data*, not merely by continuing to make our text data available for free under permissive

licenses, but by providing facilities that allow other projects to easily and productively interact with SARIT's texts; thus SARIT should be the representative, within the field of South Asian Studies, for *linked open text data*. SARIT is also poised to serve as the repository of record for high-quality text data produced either by individual scholars or by other projects. Generally, it has been members of the SARIT project who have contributed texts to the repository. In recent months, however, SARIT has received contributions from the scholarly public, guided by the documentation published on the SARIT website. SARIT may be the only resource that, on the one hand, uses TEI as its primary data format, and on the other hand, encourages contributions and corrections from the public. That this model has worked at all—and the encouraging signs that more contributions from the public are forthcoming—is testimony to the growing awareness, among scholars in South Asian Studies, that TEI documents provide a number of advantages over plain text documents. At the same time, the SARIT project should actively seek out ways to work with other projects. When we began this project in 2013, the fragmentation of digital texts projects was a major concern. Now, in 2017, it is less of a concern, given the possibility—at least in principle—of aggregating all of the text data from those separate projects. What we should strive to achieve is not hegemony, where one resource dominates the field, but interoperability, where data can be shared between resources, and those resources, which may have different strengths and different purposes, are built on a common set of standards. Greek and Roman Studies provide good examples of how the ecosystem of resources in South Asian Studies might look like in the future: there is papyri.info, which aggregates data from a number of different resources that each have their separate strengths and priorities, and the EAGLE Europeana project, which attempts to assemble all of the data concerning Greek and Latin inscriptions in Europe into a single interoperable database.

4. Bibliography and References

This bibliography includes all of the papers and presentations delivered since the beginning of the NEH-DFG funded project that concern the current and future development of SARIT. (A program of the final workshop, which took place in Vienna between May 22 and May 24, 2017, is attached to this report.)

Kellner, Birgit. 2013. "Addressing the reverse digital divide – improving Buddhological philology online." NEH-funded Conference "Advances in Digital Humanities for Buddhist Studies" (Mangalam Research Center for Buddhist Languages, Berkeley, CA, March 8-10, 2013).

———. 2016. "On Editing Sanskrit Texts Digitally – Tools, Methods and Implications." 6th Beijing International Seminar on Tibetan Studies (BISTS) (Centre for Tibetological Research, Beijing, August 1-4, 2016).

———. 2017. "Bibliography and prosopography in the digital age: EAST (Epistemology and Argumentation in South Asia and Tibet) and its challenges." Presentation delivered at "The Future of Digital Texts in South Asian Studies" (Institute for the Cultural and Intellectual History of Asia, Austrian Academy of Sciences, Vienna, May 22, 2017).

Kellner, Birgit and Liudmila Olalde. 2014. “The SARIT Project: Enriching Digital Text Collections of Buddhist Sanskrit Literature.” 17th Congress of the International Association of Buddhist Studies (University of Vienna, August 18-23, 2014).

Kellner, Birgit and Patrick McAllister. 2014. Papers on SARIT and EAST presented at the 5th International Dharmakīrti Conference (University of Heidelberg, August 26-30, 2014).

Kellner, Birgit, Patrick McAllister, Andrew Ollett, and Dominik Wujastyk. Presentations on aspects of SARIT at the workshop “Digital Visions for Indian Intellectual History: SARIT and beyond” (Institute for the Cultural and Intellectual History of Asia, Austrian Academy of Sciences, Vienna, June 30, 2016).

McAllister, Patrick. 2016. “Does this shoe fit? Applying the TEI guidelines to Sanskrit philosophical texts.” TEI Conference and Members’ Meeting 2016 (Vienna, Austria, September 30, 2016).

———. 2017. “Searching Sanskrit Texts.” Presentation delivered at “The Future of Digital Texts in South Asian Studies” (Institute for the Cultural and Intellectual History of Asia, Austrian Academy of Sciences, Vienna, May 23, 2017).

Ollett, Andrew. 2014. “*Sarit-prasāraṇam*: Developing SARIT beyond ‘Search and Retrieval.’” Buddhism and Digital Humanities Conference (University of Oxford, September 4, 2014). Slides available at https://www.academia.edu/8255307/Sarit-pras%C4%81ra%E1%B9%87am_Developing_SARIT_beyond_Search_and_Retrieval.

———. 2016. “Treasury Department: Anthologies in the Digital Age.” Digital Textualities in South Asia Conference (University of British Columbia, March 4, 2016)

———. 2017. “A Less Distant Future: Sanskrit Texts for Scholarly Communities in the Digital Age.” Presentation delivered at “The Future of Digital Texts in South Asian Studies” (Institute for the Cultural and Intellectual History of Asia, Austrian Academy of Sciences, Vienna, May 24, 2017). Slides available at https://www.academia.edu/33883303/A_Less_Distant_Future_Sanskrit_Texts_for_Scholarly_Communities_in_the_Digital_Age.

Wujastyk, Dominik. 2017. “What Do Users Want From SARIT in Future?” Presentation delivered at “The Future of Digital Texts in South Asian Studies” (Institute for the Cultural and Intellectual History of Asia, Austrian Academy of Sciences, Vienna, May 22, 2017). Slides available at https://www.academia.edu/33153494/What_Do_Users_Want_From_SARIT_in_Future.

The following are links to the most important resources developed and maintained by the SARIT project over the course of the NEH-DFG project:

- The main SARIT interface: http://sarit.indology.info.

- The “processing model” implementation of the SARIT application (still under active development): http://showcases.exist-db.org/exist/apps/Showcases/index.html (towards the bottom of the page).
- The “Simple Guidelines” for encoding Sanskrit texts: http://sarit.indology.info/exist/apps/sarit/docs/encoding-guidelines-simple.html
- The SARIT GitHub repository (for text data): https://github.com/sarit/SARIT-corpus
- The SARIT GitHub repository (for application code): https://github.com/sarit/sarit-pm

We also attach, along with this report, a spreadsheet (“e-texts in Sanskrit and Prakrit”) that contains a list of all of the texts that have been digitized by the SARIT project, including all of the texts that have been published on SARIT so far and those which are still being converted into TEI format.

This sheet includes all of the files that are currently published on SARIT or are planned to be published on SARIT. For suggestions and comments, contact the person in the "responsibility" column, or otherwise AO.

Ava	Year	Title	Author	Lang	Edition	Genre	Abbreviatic	Destinati	Respons	Status	Link	Notes
✓	2009	*Caryamelāpakapradīpa*	Āryadeva	sa	Wedemeyer, New York 2007	Tantra		SARIT	LO	updating	https://github.com/sarit/SARIT-corpus/blob/master/caryamelapakapradipa.xml	
✓	2009	*Manusmṛti*	Manu	sa	J. L. Shastri, Delhi 1983	Dharmaśāstra	manu	SARIT	LO	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/manusmrti.xml	
✓	2009	*Nāradasmṛti*	Nārada	sa	Lariviere, Philadelphia 1989	Dharmaśāstra	nāsm	SARIT	LO	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/naradasmrti.xml	
✓	2009	*Nibandhāvali*	Ratnakīrti	sa	Thakur, Patna 1975 (et varia)	Pramāṇavidyā		SARIT	PMA	updating	https://github.com/sarit/SARIT-corpus/blob/master/ratnakirti-nibandhavali.xml	
✓	2012	*Vākyapadīya*	Bhartṛhari	sa	Rau, Wiesbaden 1977	Vyākaraṇa	VāPa	SARIT	LO	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/bhartrhari-vakyapadiya.xml	
✓	2012	*Rājamārtaṇḍa*	Bhoja	sa	Manuscript owned by Suryamsu	Yoga		SARIT	LO	updating	https://github.com/sarit/SARIT-corpus/blob/master/bhoja-rajamartanda.xml	
✓	2012	*Brahmapurāṇa*	---	sa	Born digital (Tübingen Purāṇa F	Purāṇa	brapu	SARIT	LO	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/brahmapurana.xml	
✓	2012	*Mahābhārata*	Vyāsa	sa	Krishnacharya and Vyasachary	Itihāsa	MBh	SARIT		updating	https://github.com/sarit/SARIT-corpus/blob/master/mahabharata-devanagari.xml	
✓	2012	*Tattvavaiśāradī*	Vācaspati Miśra	sa	Āgāśe, Pune 1904	Yoga		SARIT	LO	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/vacaspati-tattvavaisaradi.xml	
✓	2013	*Aṣṭāṅgahṛdayasaṃhitā*	Vāgbhaṭa	sa	Kuṇṭe, Rāmacandra and Parādl	Āyurveda		SARIT	DW	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/arunadatta-sarvangasundara.xml	
✓	2013	*Sarvāṅgasundarā*	Aruṇadatta	sa	Kuṇṭe, Rāmacandra and Parādl	Āyurveda		SARIT	DW	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/arunadatta-sarvangasundara.xml	
✓	2013	*Āyurvedarasāyana*	Hemādri	sa	Kuṇṭe, Rāmacandra and Parādl	Āyurveda		SARIT	DW	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/arunadatta-sarvangasundara.xml	
✓	2013	*Aṣṭāṅgahṛdayasaṃhitā*	Vāgbhaṭa	sa	Das and Emmerick, Groningen	Āyurveda		SARIT	DW	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/astangahrdayasamhita.xml	
✓	2013	*Aṣṭāvakragītā*	---	sa	Shukla, Lucknow 1971	Vedānta	AVaGī	SARIT	LO	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/astavakragita.xml	
✓	2013	*Āyurvedasūtra*	---	sa	Shamasastry, Mysore 1922	Āyurveda		SARIT	DW	updating	https://github.com/sarit/SARIT-corpus/blob/master/ayurvedasutram.xml	
✓	2013	*Yogaśāstra*	Patañjali	sa	Āgāśe, Pune 1904	Yoga		SARIT	LO	updating	https://github.com/sarit/SARIT-corpus/blob/master/patanjalayogasastra.xml	
✓	2013	*Pramāṇavārttikālaṅkārabhā*	Prajñākaragupta	sa	Sāṅkṛtyāyana, Patna 1953	Pramāṇavidyā		SARIT	PMA	updating	https://github.com/sarit/SARIT-corpus/blob/master/pramanavarttikalankarabhasya.xml	
✓	2013	*Pramāṇavārttikavṛtti*	Manorathanandir	sa	Sāṅkṛtyāyana, Patna, 1938-194	Pramāṇavidyā	pvv	SARIT	LO	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/pramanavarttikavrtti.xml	
✓	2013	*Padārthadharmasaṅgraha*	Praśastapāda	sa	V. P. Dvivedin, Benares 1895	Vaiśeṣika		SARIT	LO	updating	https://github.com/sarit/SARIT-corpus/blob/master/prasastapada-padarthadharmasangraha.xml	
✓	2013	*Setubandha*	Pravarasena	pra	Goldschmidt, Strassburg 1880	Kāvya		SARIT	AO	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/pravarasena-setubandha.xml	
✓	2013	*Kāvyādarśa*	Daṇḍin	sa	Thakur and Jha, Darbhanga 19	Alaṅkāraśāstra		SARIT	AO	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/ratnasritika-dn.xml	
✓	2013	*Ratnaśrīṭīkā*	Ratnaśrījñāna	sa	Thakur and Jha, Darbhanga 19	Alaṅkāraśāstra		SARIT	AO	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/ratnasritika-dn.xml	
✓	2013	*Sarasvatīkaṇṭhābharaṇa*	Bhoja	sa	Kedāranāthaśarma and Vāsude	Alaṅkāraśāstra		SARIT	AO	updating	https://github.com/sarit/SARIT-corpus/blob/master/sarasvatikanthabharana-dn.xml	
✓	2013	*Rasaratnasamuccaya*	Vāgbhaṭa	sa	Aśubodhavidyābhūṣaṇa and Ni	Āyurveda		SARIT	LO	updating	https://github.com/sarit/SARIT-corpus/blob/master/vagbhata-rasaratnasamuccaya-comms.xml	
✓	2013	*Nyāyabhāṣya*	Vātsyāyana	sa	Tārānātha and Amarendramoha	Nyāya	nbh	SARIT	LO	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/vatsyayana-nyayabhasya.xml	
✓	2013	*Nyāyasūtra*		sa	Tārānātha and Amarendramoha	Nyāya	nsū	SARIT	LO	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/vatsyayana-nyayabhasya.xml	
✓	2014	*Aṣṭāṅgasaṅgraha*	Vāgbhaṭa	sa	Āṭhavale, Pune 1980	Āyurveda		SARIT	LO	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/astangasangraha.xml	
✓	2014	*Carakasaṃhitā*	Caraka	sa	Yādavaśarman, Pune 1981	Āyurveda	CaSaṃ	SARIT	DW	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/carakasamhita.xml	
✓	2014	*Kauṭalyārthaśāstra*	Kauṭalya	sa	Kangle, Bombay 1969	Arthaśāstra	Kauṭalya	SARIT	LO	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/kautalyarthasastra.xml	
✓	2014	*Pramāṇavārttika Pariśiṣṭa 1*		sa	Sāṅkṛtyāyana, Patna, 1938-194	Pramāṇavidyā		SARIT	LO	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/pramanavarttikaparisista-1.xml	
✓	2014	*Suśrutasaṃhita*	Suśruta	sa	Yādavaśarman, Bombay 1931	Āyurveda		SARIT	LO	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/susrutasamhita.xml	
✓	2014	*Tattvasaṅgraha*	Śāntarakṣita	sa	Krishnamacharya, Baroda 1926	Pramāṇavidyā	ts	SARIT	LO	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/tattvasangrahapanjika.xml	
✓	2014	*Tattvasaṅgrahapañjikā*	Kamalaśīla	sa	Krishnamacharya, Baroda 1926	Pramāṇavidyā	tsp	SARIT	LO	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/tattvasangrahapanjika.xml	
✓	2014	*Vādanyāyaṭīkā*	Śāntarakṣita	sa	Sāṅkṛtyāyana, Patna 1935-193	Pramāṇavidyā	vnt	SARIT	LO	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/vadanyayatika.xml	
✓	2015	*Nyāyabindu*	Dharmakīrti	sa	Malvania, Patna 1971	Pramāṇavidyā	nb	SARIT	LO	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/dharmottarapradipa.xml	
✓	2015	*Nyāyabinduṭīkā*	Dharmottara	sa	Malvania, Patna 1971	Pramāṇavidyā	nbṭ	SARIT	LO	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/dharmottarapradipa.xml	
✓	2015	*Dharmottarapradīpa*	Durveka Miśra	sa	Malvania, Patna 1971	Pramāṇavidyā	dp	SARIT	LO	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/dharmottarapradipa.xml	
✓	2015	*Pramāṇavārttikasvavṛttiṭīkā*	Karṇakagomin	sa	Sāṅkṛtyāyana, Allahabad 1943	Pramāṇavidyā	pvsv	SARIT	LO	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/pramanavarttikasvavrttitika.xml	
✓	2016	*Avayavanirākaraṇa*	Aśoka	sa	Thakur, Patna 1974	Pramāṇavidyā		SARIT	LO	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/avayavinirakarana.xml	
✓	2016	*Daśarūpaka*	Dhanañjaya	sa	V. Venkatacharya, Adyar 1969	Alaṅkāraśāstra		SARIT	AO	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/dasarupaka.xml	
✓	2016	*Daśarūpakāvaloka*	Dhanika	sa	V. Venkatacharya, Adyar 1969	Alaṅkāraśāstra		SARIT	AO	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/dasarupaka.xml	
✓	2016	*Daśarūpakalaghuṭīkā*	Bhaṭṭa Nṛsiṃha	sa	V. Venkatacharya, Adyar 1969	Alaṅkāraśāstra		SARIT	AO	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/dasarupaka.xml	
✓	2016	*Īśvarādimataparīkṣā*	Jitāri	sa	Bühnemann, Vienna 1982	Pramāṇavidyā	imp	SARIT	LO	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/jitari-isvaravadimatapariksa.xml	
✓	2016	*Jātinirākṛta*	Jitāri	sa	Bühnemann, Vienna 1982	Pramāṇavidyā	jāni	SARIT	LO	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/jitari-jatinirakrti.xml	
✓	2016	*Nairātmyasiddhi*	Jitāri	sa	Bühnemann, Vienna 1982	Pramāṇavidyā	nāsi	SARIT	LO	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/jitari-nairatmyasiddhi.xml	
✓	2016	*Sarvajñasiddhi*	Jitāri	sa	Bühnemann, Vienna 1982	Pramāṇavidyā	sajña	SARIT	LO	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/jitari-sarvajnasiddhi.xml	
✓	2016	*Vedāprāmāṇyasiddhi*	Jitāri	sa	Bühnemann, Vienna 1982	Pramāṇavidyā	vps	SARIT	LO	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/jitari-vedapramanyasiddhi.xml	
✓	2016	*Tantravārttika*	Kumārila Bhaṭṭa	sa	Subbāśāstrī, Pune 1929-1931	Mīmāṃsā	TaVā	SARIT	AO	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/kumarila-tantravarttika.xml	
✓	2016	*Pramāṇamīmāṃsāvṛtti*	Hemacandra	sa	Sukhlalji Sanghavi, Ahmedabac	Jaina	pramī-v	SARIT	LO	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/pramanamimamsa-and-vrtti.xml	
✓	2016	*Pramāṇamīmāṃsā*	Hemacandra	sa	Sukhlalji Sanghavi, Ahmedabac	Jaina	pramī	SARIT	LO	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/pramanamimamsa-and-vrtti.xml	
✓	2016	*Pramānāntarbhāva*	---	sa	Satīndracandra, Kolkata 1969	Nyāya	prabha	SARIT	LO	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/pramanantarbhava.xml	

This sheet includes all of the files that are currently published on SARIT or are planned to be published on SARIT. For suggestions and comments, contact the person in the "responsibility" column, or otherwise AO.

Ava	Year	Title	Author	Lang	Edition	Genre	Abbreviatic	Destinati	Respons	Status	Link	Notes
✓	2016	*Sāmānyadūṣaṇa*	Aśoka	sa	Thakur, Patna 1974	Pramāṇavidyā	sādū	SARIT	LO	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/samanyadusana.xml	
✓	2016	*Tarkabhāṣā*	Mokṣākaragupta	sa	Iyengar, Mysore 1962	Pramāṇavidyā	tabha	SARIT	LO	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/tarkabhasa.xml	
✓	2016	*Vādasthāna*	Jitāri	sa	Iyengar, Mysore 1962	Pramāṇavidyā	vastha	SARIT	LO	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/vadasthana.xml	
✓	2016	*Nyāyakaṇikā*	Vācaspati Miśra	sa	Mahaprabhulal Goswami, Bena	Mīmāṃsā	nyāka	SARIT	LO	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/vidhiviveka-and-nyayakanika.xml	
✓	2016	*Vidhiviveka*	Maṇḍana Miśra	sa	Mahaprabhulal Goswami, Bena	Mīmāṃsā	vivi	SARIT	LO	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/vidhiviveka-and-nyayakanika.xml	
✓	2016	*Rasahṛdayatantra*	Govindabhagava	sa	T.G. Kāle & D. Rasaśāstrī, Vārāṇasī 1989		RHT	SARIT	DW	updating	https://github.com/sarit/SARIT-corpus/blob/master/govindabhagavatpada-rasahrdayatantra.xml	
✓	2016	*Mugdhāvabodhinī*	Caturbhuja Miśra	sa	T.G. Kāle & D. Rasaśāstrī, Vārāṇasī 1989		MuA	SARIT	DW	updating	https://github.com/sarit/SARIT-corpus/blob/master/govindabhagavatpada-rasahrdayatantra.xml	
✓	2016	*Nyāyavārttikatātparyaṭīkā*	Vācaspati Miśra	sa	A. Thakur, Delhi 1996	Nyāya	nvt	SARIT	LO	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/nyayavarttikatatparyatika.xml	
✓	2016	*Nyāyasūtra*		sa	A. Thakur, Delhi 1996	Nyāya	nsū	SARIT	LO	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/nyayavarttikatatparyatika.xml	
	2016	*Saṃmatitarkaprakaraṇam*	Siddhasena Divā	pra	Sukhlalji Sanghavi and Becharc	Jaina	stp	SARIT	LO	Up to date		No full text access.
	2016	*Tattvabodhavidhāyinī*	Abhayadevasūri	sa	Sukhlalji Sanghavi and Becharc	Jaina	tbv	SARIT	LO	Up to date		No full text access.
✓	2016	*Tarkarahasya*	---	sa	Yaita, Narita 2005	Pramāṇavidyā	tara	SARIT	LO	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/tarkarahasya.xml	
✓	2017	*Siddhiviniścaya*	Akalaṅka	sa	Mahendrakumār, Vārāṇasī 195	Jaina	svi	SARIT	LO	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/siddhiviniscayatika.xml	
✓	2017	*Siddhiviniścayavṛtti*	Akalaṅka	sa	Mahendrakumār, Vārāṇasī 195	Jaina	svi-v	SARIT	LO	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/siddhiviniscayatika.xml	
✓	2017	*Siddhiviniścayaṭikā*	Anantavīrya	sa	Mahendrakumār, Vārāṇasī 195	Jaina	siviṭ	SARIT	LO	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/siddhiviniscayatika.xml	
✓	2017	*Dvādaśāraṃ Nayacakram*	Mallavādi	sa	Jambūvijaya, Bhavnagar 1966	Jaina	nayacakra	SARIT	LO	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/dvadasara_nayacakra.xml	
✓	2017	*Nyāyāgamānusāriṇī*	Siṃhasūri	sa	Jambūvijaya, Bhavnagar 1966	Jaina		SARIT	LO	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/dvadasara_nayacakra.xml	
✓	2017	*Nyāyamañjarī*	Jayanta Bhaṭṭa	sa	Varadacharya, Mysore 1969	Nyāya	nyāma	SARIT	LO	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/nyayamanjari.xml	
✓	2017	*Nyāyasaurabhaṭippaṇī*	K. S. Varadachar	sa	Varadacharya, Mysore 1969	Nyāya		SARIT	LO	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/nyayamanjari.xml	
✓	2017	*Nyāyasūtra*		sa	Varadacharya, Mysore 1969	Nyāya	nsū	SARIT	LO	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/nyayamanjari.xml	
✓	2017	*Nyāyakumudacandra*	Prabhācandra	sa	Mahendrakumār, Bombay, 193	Jaina	nkm	SARIT	LO	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/nyayakumudacandra.xml	
✓	2017	*Laghīyastraya*	Bhaṭṭākalaṅkade	sa	Mahendrakumār, Bombay, 193	Jaina	laghī	SARIT	LO	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/nyayakumudacandra.xml	
✓	2017	*Laghīyastrayavivṛti*	Bhaṭṭākalaṅkade	sa	Mahendrakumār, Bombay, 193	Jaina	laghī-v	SARIT	LO	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/nyayakumudacandra.xml	
✓	2017	*Dharmottaraṭippaṇaka*	---	sa	Yaita, Narita 2005	Pramāṇavidyā	dhṭ	SARIT	LO	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/dharmottaratippanaka.xml	
✓	2017	*Nyāyasudhā*	Someśvara Bhaṭ	sa	Mukunda Śāstrī, Benares, 1909	Mīmāṃsā		SARIT	AO	Up to date	https://github.com/sarit/SARIT-corpus/blob/master/nyayasudha.xml	
	2017	*Rasayanakhanda of the Ras*	Nityanātha Siddh	sa	Yādavaśarmā Trivikrama Ācārya & Rāmacandraśāstrī Paṇaśī			SARIT	DW	Up to date		
	2017	*Kalyanakaraka*	Ugraditya	sa				SARIT	DW	conversion		
✓	2017	*Yogapradipa*		sa	MS Chinmaya International Foundation (CIF), EAP729/1/2/66(			SARIT	DW	uptdating	https://github.com/sarit/SARIT-corpus/blob/master/yogapradipa.xml	
✓	2017	*Saduktikarnamrta*	Śrīdharadāsa	sa	Sures Chandra Banerji, Calcutta, 1965			SARIT	DW	updating	https://github.com/sarit/SARIT-corpus/blob/master/sridharadasa-saduktikarnamrta.xml	
		Jñānaśrīmitranibandhāvali	Jñānaśrīmitra	sa	A. Thakur, Patna 1987	Pramāṇavidyā		SARIT	PMA	conversion		
		Bṛhatī	Prabhākara Miśr	sa	Ramanatha Sastri, Madras, 193	Mīmāṃsā		SARIT	AO	conversion		
		Ṛjuvimalā Pañcikā	Śālikanātha Miśr	sa	Ramanatha Sastri, Madras, 193	Mīmāṃsā		SARIT	AO	conversion		
		Śṛṅgāraprakāśa	Bhoja	sa	R. Dvivedī, Delhi 2007	Alaṅkāraśāstra		SARIT	AO	Copyright limbo		
		Nāṭyaśāstra	Bharata	sa	G.O.S. Edition (Baroda 1926-2(	Nāṭyaśāstra		SARIT	AO	conversion		
		Abhinavabhāratī	Abhinavagupta	sa	G.O.S. Edition (Baroda 1926-2(	Nāṭyaśāstra		SARIT	AO	conversion		
		Kāvyamīmāṃsā	Rājaśekhara	sa	Dalal and Sastry, Baroda 1934	Alaṅkāraśāstra		SARIT	AO	planned		
		Kāvyānuśāsana	Hemacandra	sa	Parikh, Bombay 1938	Alaṅkāraśāstra		SARIT	AO	planned		
		Dhvanyālokalocana	Abhinavagupta	sa		Alaṅkāraśāstra		SARIT	AO	planned		
		Vyaktiviveka	Mahimabhaṭṭa	sa		Alaṅkāraśāstra		SARIT	AO	planned		
		Abhidhāvṛttamātṛkā	Mukulabhaṭṭa	sa		Alaṅkāraśāstra		SARIT	AO	planned		
		Ṭupṭīkā	Kumārila Bhaṭṭa	sa		Mīmāṃsā		SARIT	AO	planned		
		Prakaraṇapañcikā	Śālikanātha Miśr	sa	Subrahmaṇyaśāstri, Vārāṇasī 1	Mīmāṃsā		SARIT	AO	planned		
		Tattvabindu	Vācaspati Miśra	sa	Rāmasvāmiśāstrī, Trichinopoly	Mīmāṃsā		SARIT	AO	planned		
		Śāstradīpikā	Pārthasārathi Mi	sa	Dharmadattasūri, Bombay 1915	Mīmāṃsā		SARIT	AO	planned		
		Vidhirasāyana	Appayya Dīkṣita	sa	Subrahmaṇyaśāstrī, Vārāṇasī 1	Mīmāṃsā		SARIT	AO	planned		
		Śivārkamaṇidīpikā	Appayya Dīkṣita	sa	Halasyanath Shastri, 1908	Vedānta		SARIT	AO	conversion		

The Future of Digital Texts in South Asian Studies — A SARIT Workshop

2017-05-22 to *2017-05-24*

http://www.ikga.oeaw.ac.at/Events/SARIT_Workshop_2017

Birgit Kellner, Patrick McAllister, Andrew Ollett

Version: *2017-05-20*

Contents

1 Introduction and practical information

As a conclusion to a four-year project dedicated to developing and enriching a collection of digital texts in Sanskrit and other Indian languages, the team behind SARIT is convening a workshop called "The Future of Digital Texts in South Asian Studies." The goal is twofold. First, we want to survey and reflect on the current state of digital texts in our field. What is a "digital text"? How are they produced? Who is responsible for them? How are they provided to users? Who are their users, and what do they do with them? How, if at all, have they changed the landscape of research and teaching? Second, we want to reflect on the future of digital texts. What could we be doing with them that we aren't doing yet? What inspiration can we take from projects in other fields? What emerging technologies can we take advantage of? How can we better integrate our various digital projects? How can we involve communities of students, teachers, and researchers in the production, curation, and publication of digital texts?

The workshop is also intended to stimulate discussion on the future of SARIT.

1.1 Practical Information

- Where:
  - 'Seminarraum' (room 2.25), Institute for the Cultural and Intellectual History of Asia (IKGA)
  - Hollandstraße 11+13/2nd floor, 1020 Vienna, Austria. (map)
- When: *2017-05-22* to *2017-05-24*
- Contacts:
  - patrick.mcallister@oeaw.ac.at
  - office.ikga@oeaw.ac.at
  - T: (+43 1) 515 81 / 6400
- Registration: to help us prepare, please register per email to **both** contacts, office.ikga@oeaw.ac.at and patrick.mcallister@oeaw.ac.at no later than *2017-05-07*.

2 Schedule

2.1 *2017-05-22*

Registration will be open from **9:00** in the 'Sekretariat', room 2.49, at the IKGA.

Opening session

1. **10:00–10:30** Birgit Kellner: *The development of SARIT 2013–2017: goals, achievements, problems*

2. **10:30–11:00** Dominik Wujastyk: *What do users want from SARIT in future?*
3. **11:00–11:30** Coffee break

History through Indic Texts (1) // Chair: Masahiro Shimoda

1. **11:30–12:30** Bronner, Yigal: *Indic Prosopography in the Digital Age*
2. **12:30–14:00** Lunch break
3. **14:00–15:00** Kellner, Birgit: *Bibliography and prosopography in the digital age: EAST (Epistemology and Argumentation in South Asia and Tibet) and its challenges*
4. **15:00–16:00** Baums, Stefan: *Documents, Databases and Networks: Scholarly Work on Gāndhārī in the Digital Age*
5. **16:00–16:30** Coffee break
6. **16:30–18:30** Software Demonstrations

2.2 *2017-05-23*

History through Indic Texts (2) // Chair: Nina Mirnig

1. **10:00–11:00** Bajracharya, Manik and Christof Zotter: *Turning pre-modern documents into digital texts: The pragmatics of an approach*
2. **11:00–11:30** Coffee break

Computational Linguistics for Indic texts // Chair: Lou Burnard

1. **11:30–12:30** Scharf, Peter M.: *Creative and intelligent use of linguistic, textual, and bibliographic information to enhance interlinked access to lexical, textual, and image data*
2. **12:30–14:00** Lunch break
3. **14:00–15:00** Kulkarni, Amba: *Bridging the gap between Computational tools and Sanskrit Digital Libraries: Where do we stand?*
4. **15:00–16:00** McAllister, Patrick: *Searching Sanskrit Texts*
5. **16:00–16:30** Coffee break

Computer-assisted Editing of Indic Texts (1)

1. **16:30–17:30** Bellefleur, Tim and Adheesh Sathaye: *Developing Linked Data Standards for Working with Sanskrit Manuscript Traditions*
2. **17:30–18:30** Balogh, Dániel: *Building a Database of Indic Inscriptions*
3. **19:00–22:00** Dinner

2.3 *2017-05-24*

Computer-assisted Editing of Indic Texts (2) // Chair: Karlheinz Mörth

1. **10:00–11:00** Li, Charles: *Editors as Maintainers*
2. **11:00–11:30** Coffee break
3. **11:30–12:30** Tomabechi, Toru: *TEI Markup of Abhayākaragupta's Āmnāya-mañjarī: An Attempt to Create an "Open Research Note" for the Study of Late Indian Buddhism*
4. **12:30–14:00** Lunch break
5. **14:00–15:00** Maas, Philipp: *Sanskrit Textual Criticism in the Digital Age – Will Really Everything Change?*
6. **15:00–16:00** Ollett, Andrew: *A Less Distant Future: Sanskrit Texts for Scholarly Communities in the Digital Age*
7. **16:00–16:30** Coffee break

Closing

1. **16:30–18:00** Closing words, round-table discussion: *What next for SARIT?*

3 Participants and Abstracts

3.1 Balogh, Dániel

- British Museum
- danbalogh@gmail.com

Building a Database of Indic Inscriptions

This paper introduces a recent initiative in digital epigraphy under the aegis of the ERC Synergy project 'Beyond Boundaries – Religion, Region, Language and the State.' The project as a whole aims to re-evaluate the social and cultural history of the Gupta period in South, Central and Southeast Asia and approach an understanding of the region as an interconnected cultural network. One component of this project is the 'Siddham' database of Indic epigraphic texts. Its development was commenced in the summer of 2015 with the encoding of previously published Sanskrit inscriptions created under the imperial Gupta rulers. It will be progressively expanded both horizontally (by adding inscriptions from other dynasties and regions) and vertically (by accumulating metadata, gradually increasing the granularity of markup, and through re-editing crucial inscriptions).

EpiDoc (an application of TEI for encoding epigraphic documents) serves as the flesh and blood of our corpus: texts are stored in XML snippets, each

comprising the edition division of a full EpiDoc file. Siddham's skeleton is made up of relational database tables. The edition snippets, along with other snippets containing translations (and, optionally, critical apparatus and commentaries), are referenced from an "Inscriptions Table" that additionally stores metadata pertinent to each inscription, such as layout and hand description, language and date. A separate "Objects Table" serves as the repository of metadata pertaining to inscription-bearing objects, such as physical properties (material, dimensions and freeform description) and history. The separation of object metadata from inscription metadata is conceptually desirable as it brings objects to the fore as entities in their own right rather than mere dismissible substrates of the texts they carry. It is our hope that Siddham will not only become a useful reference tool for textual scholars of Indic languages, but will, through the foregrounding of the objects themselves and through the inclusion of translations, also encourage the formulation of new types of questions that scholars of various disciplines may ask of text-bearing objects of this region.

3.2 Baums, Stefan

- Bavarian Academy of Sciences and Humanities, Ludwig Maximilian University of Munich
- baums@lmu.de

Documents, Databases and Networks: Scholarly Work on Gāndhārī in the Digital Age

Gāndhārī is a Middle Indo-Aryan language attested from the third century BCE until the fifth century CE. Until quite recently, it was known almost exclusively from inscriptions, administrative documents and a single literary manuscript, and this scarcity and specialization of sources meant that no comprehensive dictionary or grammar of the Gāndhārī language were ever attempted. The situation changed radically with the discovery, in the 1990s, of about one hundred long manuscripts containing Buddhist and literary texts, which are now in the process of being edited. The new manuscript discoveries prompted Andrew Glass and myself to undertake (in the year 2002) the compilation of a digital corpus of Gāndhārī texts as the basis for our Dictionary of Gāndhārī. Our corpus reached completion several years ago (with a total of currently 2,751 texts), and we have made significant progress in lemmatization, morphological marking and article writing (currently 6,713 articles covering 33,403 references) with incipient support for syntactic treebanking. We make our complete Gāndhārī corpus as well as our in-progress lexicographic work available from our website http://gandhari.org (which also provides a selection of Old and Middle Indo-Aryan dictionaries for the convenience of users).

In parallel with our lexicographic work, the field of Gāndhārī manuscript studies has broadened significantly in the last twenty years, with two major projects (the Early Buddhist Manuscripts Project in Seattle and the Buddhist Manuscripts from Gandhāra Project in Munich) now involved in the edition of the new discoveries. To support the work of these two projects in particular, and of scholars of early South Asian documents more generally, we spearheaded the development of a new digital toolset called Research Environment for Ancient Documents (READ) that has received funding and adoption from a number of institutions. Design decisions for READ were shaped by an analysis of the workflows of the target user communities and of the intended scholarly products, as well as by social considerations such as the representation of individual scholarly work and distinct project identities. Two defining characteristics of READ are the linked parallel storage of multiple editions of the same source text, and the linking of images and transcriptions as a basis for navigation, paleographic work and pedagogical applications. The core technologies behind READ are relational databases (PostgreSQL) for storage, a PHP/HTML/JavaScript system for interaction with stored content and for content creation, and TEI P5 XML documents for export, import and archival purposes. Capabilities for networking textual corpora and analytical works on different servers (either running READ or providing other TEI-based interfaces) are currently under development. At the same time as this technical work, the editors of the Gandhāran Buddhist Texts series made new publication and archival arrangements ensuring the perpetual availability of editions in the series under open-access licensing and in open data formats, both online and through print on demand.

The present paper will give an overview of the history and prospects of these digital initiatives for Gāndhārī studies and related fields. It will consider some technical differences between relational and document databases and how they relate to the traditional organization of scholarly work. It will argue that the combination of open formats and licenses with a document level of organization and networking capabilities is key to ensuring a proper balance between content sharing on the one hand and local resource management and the integrity of individual scholarly work on the other.

3.3 Bajracharya, Manik and Christof Zotter

- manik.bajracharya@adw.uni-heidelberg.de
- christof.zotter@adw.uni-heidelberg.de
- Documents on the History of Religion and Law of Pre-modern Nepal, Heidelberg Academy of Science and Humanities

Turning pre-modern documents into digital texts: The pragmatics of an approach

Digital humanities provide powerful tools that can facilitate the research work and enhance the options of scientific analysis, also in South Asian studies. But, having their own requirements, digital solutions can also consume a lot of man power, especially if they need to be tailored to the needs of researchers and users. Furthermore, it requires "esoteric" technical knowledge to understand what is possible and how it can be realised; a knowledge that is not a usual part of the curriculum in South Asian Studies. With a special focus on the production of digital texts, this contribution will report about the experiences of a humanities' project that decided to enter the challenges of working digitally. It will address some of the questions that accompanied the process of setting up, extending and refining the project's IT structure and will give an account of a still ongoing quest for workable solutions, the cost-benefit-ratio in defining standards, the organisation of workflows, collaborative work and authorship, and the unavoidable compromises.

The research unit "Documents on the History of Religion and Law of Premodern Nepal" (see http://www.haw.uni-heidelberg.de/forschung/forschungsstellen/nepal/welcome.html) of the Heidelberg Academy of Science and Humanities employs digital humanities in various ways. It has built up a MySQL database to enter catalogue data of a huge corpus of historical documents in order to make them accessible to the project members but also to the general public, and it provides digital texts, namely editions of selected documents prepared according to the standards of the Text Encoding Initiative (TEI). These two building blocks are linked with each other and associated with a bibliography, a glossary and, as latest feature, an ontology of named entities. The planning and implementation of all these components went along with sometimes extensive discussions about the issues mentioned above. Similar processes will accompany future steps of the project, such as the involvement of a lemmatizer, OCR and the automatized annotation of at least some formal features.

One big advantage of digital texts is that associated information is extensible and that they allow for different reuses. However, to keep a text 'alive' it needs not only convertible formats and a manageable technical habitat, but also an, at least partly, new understanding of what is actually being produced.

3.4 Bellefleur, Tim and Adheesh Sathaye

- tbelle@alumni.ubc.ca
- University of British Columbia
- Adheesh.Sathaye@ubc.ca
- Dept. of Asian Studies, University of British Columbia

Developing Linked Data Standards for Working with Sanskrit Manuscript Traditions

In the world of digital textual scholarship, we generally focus either on creating artifacts (e-texts, collations, stemmata) or applications—the software tools that create, process, and present these artifacts. While SARIT has, quite rightly, focused its efforts on developing and maintaining encoding standards for artifacts—especially e-texts—the needs of modern digital philology require a more robust set of standards for how different applications may interface with such artifacts, as well as with one another. The development of such standards, we suggest, may help us get one step closer to the "holy grail" of an extensible and user-friendly digital environment for Sanskrit textual scholarship. In this joint presentation, we will first present the perspective of the end-user, the textual scholar, and explore why one would need networked textual data that is streamlined, multi-dimensional, and responsive to human decision-making and workflows. We will then offer some solutions that we have been developing as part of the Vetāla Project at UBC through the use of Linked Data standards, such as Open Annotations, to facilitate the robust connections between artifacts and applications.

3.5 Bronner, Yigal

- The Hebrew University of Jerusalem
- yigal.bronner@mail.huji.ac.il

Indic Prosopography in the Digital Age

Panditproject.org is a digital humanities project with a unique and ambitious task: to create a database for the vast world of South Asian letters. The name stands for the Sanskrit title of a virtuoso scholar with full mastery of traditional knowledge systems, but as an acronym it also expresses the project's main objective: the creation of a Prosopographical Database of Indic Texts. In brief, Panditproject.org seeks to store, curate, and share reliable data on works, people, places, institutions, and manuscripts from premodern South Asia, in addition to relevant secondary sources, and to do so across period, language, discipline, and subject matter. It is designed as an interactive web-based repository that scholars of every South Asian specialty and interest can contribute to and as a basic tool on which they will routinely come to rely. I propose to give a hands-on presentation of the database and its possibilities and remaining challenges.

3.6 Burnard, Lou

- lou.burnard@retired.ox.ac.uk

- Former Assistant Director of Oxford University Computing Services (OUCS)
- Central contributor to the Text Encoding Initiative (TEI)

Lou Burnard will chair the panel Computational Linguistics for Indic texts.

3.7 Hellwig, Oliver (*Cancelled*)

Unfortunately, this presentation had to be cancelled.

- SFB 991, University of Düsseldorf
- oliver.hellwig@indsenz.com

Machine Learning Techniques in an Indological Context

Although Digital Humanities as a research paradigm have promoted the interaction between text-oriented Philology on one and quantitative Natural Language Processing (NLP) and Machine Learning (ML) on the other side, there is still an enormous conceptual gap between these two approaches to texts and language. In most research scenarios, neither side is aware of specific problems and solutions presented by the other one. While NLP and ML provide efficient frameworks for learning and for reasoning based on partly observable information, philological disciplines assemble the specialized knowledge and the – partly undigitized – resources for understanding historical texts. With the exception of few "fashionable" areas such as topic or graph analysis, Digital Humanities as the "boundary discipline" has not been able to connect both fields effectively.

This paradigmatic separation has considerable consequences. Most NLP studies work with the same datasets that cover a limited set of modern languages[1] and of intellectual domains (mainly newspaper texts). They often silently assume that mechanisms working for the benchmark data sets of modern newspaper English will be equally efficient on out-of-domain texts and/or texts in other (ancient) languages. Philology, on the other hand, frequently does not make use of standard methods from NLP or ML that could be helpful in structuring the available data and in drawing scientifically sound conclusions from them.

The presentation will focus on two quantitative approaches that can strongly increase the efficiency of philological reasoning.

1. **Supervised classification** deals with predicting the class of a new instance, given a set of previously labeled instances. Supervised classification is especially useful in the context of corpus annotation, where

1. English and Chinese; modern German, for example, is almost considered as an under-resourced language.

new instances (e.g., unannotated words or syntactic structures) should be labeled automatically by using an ML model. The presentation will introduce several research cases, in which **Deep Learning** models are used for the morphological, lexical, semantic, and syntactic annotation of Sanskrit texts.

2. Although Topic Models are quite popular in Literary Studies, the underlying field of **Graphical Models** has not found much attention in philological research. Graphical models provide a principled method of evaluating causal relationships between large numbers of variables in textual data, and allow to draw conclusions about "hidden factors" such as authorship given only a limited set of observed data (words, topics). The presentation will given an informal introduction into the theory underlying Graphical Models, and will sketch how problems of authorship attribution and text stratification can be formulated in this framework.

Useful Links

- *ML* Graphically appealing and non-technical introduction into Machine Learning: http://www.r2d3.us/visual-intro-to-machine-learning-part-1/
- *Graphical Models* A comparatively non-technical introduction to Graphical Models, with some interesting pointers to phylogenetics and document processing: https://projecteuclid.org/download/pdfview_1/euclid.ss/1089808279
- *Deep Learning* The "Deep Learning Tsunami" and Computational Linguistics: http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00239

3.8 Kellner, Birgit

- Institute for the Cultural and Intellectual History of Asia, Austrian Academy of Sciences
- birgit.kellner@oeaw.ac.at

Bibliography and prosopography in the digital age: EAST (Epistemology and Argumentation in South Asia and Tibet) and its challenges

In 1995, Ernst Steinkellner's and Michael Torsten Much's systematic survey of the literature of the logico-epistemological school of Buddhism, chiefly in India, was published in print, as part of a larger endeavour to systematically document Buddhist Sanskrit Literature. (Systematischer Überblick über die Literatur der erkenntnistheoretisch-logischen Schule des Buddhismus. Göttingen 1995: Vandenhoeck & Ruprecht; Systematische Übersicht über die buddhistische Sanskrit-Literatur 2).

The survey offered whatever biographical information was available at the time about individual thinkers chiefly of the *pramāṇa* tradition. But its main goal was to comprehensively document publications in a well-defined system informed by the textual scholar's main interests, using works as the main anchor of classification. Publications were categorized in terms of whether they contained full and partial editions or translations, whether they offered textual fragments, or contained glossaries and indices.

Efforts to transform the data provided by Steinkellner and Much into a database structure date back to the early 2000s (with financial support by the Austrian Science Fund FWF), but it was only within the framework of, first, the Cluster of Excellence "Asia and Europe in a Global Context" of the University of Heidelberg and, then, the DFG-NEH supported SARIT project that a major push could be made to produce a new interactive digital resource pursuing the same overarching scholarly goal – comprehensive documentation of logico-epistemological literature – while making use of new technological possibilities to continually update information and offer it in a form that was better attuned to the dynamics of the Web. The resource EAST (Epistemology and Argumentation in South Asia and Tibet, http://east.uni-hd.de) was released in 2011, and has been updated ever since.

In this contribution I shall offer a brief presentation of EAST, but mainly use EAST as an example for a more general discussion of how the move from print to digital offers challenges for the production and maintenance of prosopographical and biographical resources (with particular focus on South Asian Studies).

3.9 Kulkarni, Amba

- Indian Institute of Advanced Study, Shimla
- ambapradeep@gmail.com

Bridging the gap between Computational tools and Sanskrit Digital Libraries: Where do we stand?

The last decade has witnessed vibrant activities in the field of Sanskrit Computational Linguistics. Several tools have been developed performing various tasks such as word analysis, word generation, segmenting a sandhied text into meaningful components, compound analysis and dependency parsing. Unlike other natural languages, Sanskrit has the advantage of having an almost exhaustive grammar. At the same time it is unique in allowing parallel meanings running across the texts. Both these features pose challenges for a computational linguist, since now there is a demand to produce all possible analyses in parallel by providing the justification in terms of grammar rules.

In this paper I describe the efforts in building **Saṁsādhanī**, a platform for Sanskrit Computational Linguistics that has the features described above.

The platform operates interactively with the **Heritage** segmenter to produce various possible analyses. The interactive user interface is developed to share the load between man and machine in such a way that tasks hard for human being are done by the machine and vice versa.

This platform is being used to develop e-readers for various Classical texts such as Śrīmad Bagavadgītā, Śiśupāvadham, Bhaṭṭikāvya, Mahābhārata, etc. serving dual purpose. On the one hand they demonstrate the effective use of technology in reviving the traditional methods of teaching following the Khanḍānvaya and on the other hand they help in generating annotated texts that will help bootstrapping the Machine learning efforts.

3.10 Li, Charles

- University of Cambridge
- cchl2@cam.ac.uk

Editors as Maintainers

The emergence of electronic texts, and our increasing reliance upon them, has made the shortcomings of printed editions readily apparent. However, most electronic texts are still conceived of as digital facsimiles of printed books; they are not authoritative in their own right, and they usually strive only to reproduce the printed text faithfully, even if that text might be incorrect. There is no framework in which corrections to published editions can be suggested by third parties, or in which new evidence can be incorporated into an existing edition, other than by re-editing it and re-publishing it. In order to change this, we would need to change our conception both of the edition itself — not as a fixed text, but as a body of evidence and hypotheses that are progressively improved — and of the role of an editor — not as a compiler of a fixed text, but as a maintainer of a textual tradition, who takes on the task of reviewing suggestions, corrections, and new evidence, and updates the edition as necessary. In this model, an edition would have one or more maintainers, who oversee the project, and a virtually unlimited number of contributors, whose work — in the form of transcriptions, emendations, testimonia, critical notes, etc. — would be recorded with a version control system.

Technology Demonstration

In the process of preparing a new critical edition of Bhartṛhari's Dravyasamuddeśa with the commentary of Helārāja, I have developed some open source tools to collate diplomatic transcripts of manuscript witnesses and to display an interactive apparatus alongside the text. Compared to a traditional, printed edition, this digital edition does not treat the apparatus as part of the

content, but as a dynamically-generated analysis of the content, which consists of the witnesses themselves. Each witness is considered a text in its own right, and each transcript a faithful representation of that text. In this way, the edition becomes a collection of documents that can be easily added to if new witnesses are discovered. I will be demonstrating the user interface of the edition both from the point of the view of an editor — covering the transcription process and methods of collaboration —, and from the point of view of a reader — exploring the interactive tools for researching textual variation. The online edition can be found at http://saktumiva.org/wiki:dravyasamuddesa:start, and the source code for the software is on GitHub (http://github.com/chchch/upama).

3.11 Maas, Philipp

- University of Leipzig
- Philipp.A.Maas@gmail.com

Sanskrit Textual Criticism in the Digital Age – Will Really Everything Change?

Indology and South Asian Studies research the cultures of South Asia in their historical contexts. To this end, these disciplines strongly depend on information contained in primary sources written in Sanskrit and other languages, among which the works of the multiple genres of literature are highly important. However, most works of Sanskrit literature are no longer available in the version in which they were originally composed and written down. All that is available are printed editions based on mostly unidentified manuscript copies. Moreover, approximately six million mostly unsought manuscripts exist, which are mostly copies produced from previous copies along unknown (but not unknowable) lines of transmission. This manuscripts heritage, the largest of all cultures worldwide, is the object of research of Sanskrit textual criticism. Its double interrelated objective is traditionally regarded as (1) reconstructing as exactly as possible a text version that resembles as closely as possible a text version that would have been acceptable to the author or final redactor of a given work, and (2) investigating the transmission history of the work under research.

In spite of the fundamental importance of critical editing for any kind of text based research, only very few critical editions of Sanskrit works have been produced so far. And virtually all existing critical editions were designed to be published in print. Now, the so-called digital revolution provides scholars with completely new options and possibilities for critically editing and for the publication of their work. Peter Robinson, a pioneer and leading expert in digital editing works of English literature, has recently argued that new technical tools and web-based publication facilities may

fundamentally change the methods and aims of critically editing (P. Robinson, "The Digital Revolution in Scholarly Editing." Eds. Barbara Crostini et al., Ars Edendi Lecture Series. Vol. 4. Stockholm 2016, pp. 181–201 [http://www.stockholmuniversitypress.se/site/books/10.16993/baj/]). The present talk will critically examine Robinson's assessments with special reference to the aims and objectives for Sanskrit textual criticism as it is (and will be) applied in the DFG-sponsored project "A Digital Critical Edition of the Nyāyabhāṣya" at the University of Leipzig, Germany.

3.12 McAllister, Patrick

- Institute for the Cultural and Intellectual History of Asia, Austrian Academy of Sciences
- patrick.mcallister@oeaw.ac.at

Searching Sanskrit Texts

Searching electronic data is a large and complex field in modern information technology. Simple search interfaces belie the sophistication of both the theoretical foundation and the practical engineering that make these searches possible. Over the last decades, important parts of this technology, on the theoretical as well as on the practical side, have become publicly available. This has made it possible to apply these tools to the searching of texts in languages that are decidedly not in the center of interest for most search companies and also most professional programmers, such as Sanskrit.

This paper will first give an overview of the search techniques most commonly used for Sanskrit texts, with a specific focus on the ones used on SARIT's public interface (http://sarit.indology.info), arguably the most advanced public search engine for Sanskrit texts, along with their respective advantages and limitations. After that, several less used search methods will be investigated as to their utility for searching Sanskrit texts, particularly ones based on synonyms, phonetic algorithms, and translations. From an evaluation of these various approaches, guided by some practical considerations, the reasonable expectations for future improvements to the computer-assissted searching of Sanskrit texts will become clearer.

3.13 Mirnig, Nina

- Nina.Mirnig@oeaw.ac.at
- Institute for the Cultural and Intellectual History of Asia, Austrian Academy of Sciences

3.14 Mörth, Karlheinz

- Karlheinz.Moerth@oeaw.ac.at
- Director of the Austrian Centre for Digital Humanities (ACDH-OeAW)
- Austrian Academy of Sciences

Dr. Karlheinz Mörth will chair the panel Computer-assisted Editing of Indic Texts (2).

3.15 Ollett, Andrew

- andrew.ollett@gmail.com
- Harvard University

A Less Distant Future: Sanskrit Texts for Scholarly Communities in the Digital Age

In the current funding cycle for SARIT, the Columbia University subproject has prepared a series of texts, with a focus on poetics (*alaṃkāraśāstra*) and hermeneutics (*mīmāṃsā*). At the start of the project, we were relatively new to TEI, and believed that it could improve on the existing models, both formal and informal, of how people interact with texts in our field. First, TEI texts have an advantage over printed texts in that their structure and content is machine-readable. For most users, this simply means "searchable," but we were interested in a wide range of other possible applications: named entity recognition, alignment, identification of quotations, word cooccurrence patterns, and so on. Second, TEI can represent features of a printed edition that plain text files typically don't, including notes, front and back matter, a critical apparatus, pagination and lineation, and so on. Our hope was that, by putting all of this information into the digital text, the digital text would be as "citeable" for scholarly purposes as the printed edition on which it was based. That is, in addition to being "machine-readable," the text would be "scholar-readable." Third, we hoped that our texts would be dynamic rather than static, open to the scholarly community for further improvement and annotation. Our test-case would have been Abhinavagupta's *New Dramatic Art* (*Abhinavabhāratī*), in which a careful reader can conjecturally improve the printed edition on almost every single page. These texts should therefore also be "community-readable." This talk will cover the progress we've made, and the challenges we've faced, in producing machine-, scholar-, and community-readable texts. We have found that there are two limiting reagents in this process: the considerable human labor involved in converting printed editions to high-quality TEI texts (which grows exponentially when, as is often the case in our field, the typographic and editorial conventions of the source edition are inconsistent), and the total inadequacy of existing applications for interacting with these kinds of texts. How can these limitations be addressed or

overcome? We'll share some suggestions from our experience, and from new approaches to TEI publishing that SARIT is now taking advantage of. We'll also offer a model for interacting with digital texts—reading with a selection of commentaries, facsimiles of printed editions and manuscripts, navigable cross-references, a critical apparatus, annotations, and bibliographic information at one's fingertips—that is much closer to realization now than it was when this project started.

3.16 Sathaye, Adheesh and Tim Bellefleur

- Adheesh.Sathaye@ubc.ca
- Dept. of Asian Studies, University of British Columbia

See Bellefleur, Tim and Adheesh Sathaye for the abstract.

3.17 Scharf, Peter M.

- scharfpm7@gmail.com
- President, The Sanskrit Library; Visiting Professor, IIT Bombay

Creative and intelligent use of linguistic, textual, and bibliographic information to enhance interlinked access to lexical, textual, and image data

It is obvious to participants in the SARIT workshop on the future of digital texts in South Asian studies that we are in the midst of a media transition. Just as there was a transition in the primary mode of transmission of knowledge from oral to written, and from written to printed, we are now undergoing a transition from the printed to digital medium. What is not obvious is how the new digital medium liberates us from the conventions appropriate for the written and print media dictated primarily by visual factors, and how to maximize the potentialities of the digital medium by utilizing linguistic, text-structural, and bibliographic information. Clarification of encoding principles resolves on linguistically precise character encoding for functions such as searching, morphological identification, and parsing. Clear delineation of data-entry and display functions from linguistic processing grants freedom of these functions to conform to human efficiency considerations and user preferences. Judicious use of Text-Encoding Initiative (TEI) markup likewise separates XML text markup from optional display formats and permits interlinking of text with lexical, linguistic, related textual and bibliographic resources on the one hand, and with images on the other. The integration of image analysis software, such as OCR software, with textual and bibliographic information permits the development of approximate image finding

aids by automated methods, and precise finding aids by including human supervision.

The Sanskrit Library (http://sanskritlibrary.org) has developed linguistically precise phonetic encodings for Sanskrit, revised the Unicode Standard to include Vedic characters, and developed comprehensive transcoding software for interchange between linguistic processing, data-entry encodings, and standard Romanization or Indic script Unicode display. Texts are linked to morphological analysis software, and digital lexical sources, and in exemplary distributed collaboration, with the Sanskrit Heritage parser. An integrated dictionary interface permits lookup in some forty different lexical sources including the major bi-lingual and monolingual dictionaries as well as little known and under-utilized specialized dictionaries. The Sanskrit Library has also created a pipeline for digital cataloguing of Sanskrit manuscripts including a template that incorporates the standards of the American Committee for South Asian Manuscripts (ACSAM) and conforms to TEI manuscript guidelines. Manuscript passages are optionally displayed in text-structure or manuscript format and are linked to corresponding searchable digital texts and to manuscript images. A comprehensive catalogue currently contains 160 entries for Sanskrit manuscripts at Brown University and corresponding manuscripts at the University of Pennsylvania and 1800 draft entries for Sanskrit manuscripts at Harvard University. A manual interface allowed association of passages in manuscript images with corresponding digital texts for the Brown and Penn mss. Newly developed text-image alignment software utilizes dirty OCR along with textual and bibliographic parameters supplied by catalogue entries to estimate the location of passages automatically with surprising accuracy.

3.18 Shimoda, Masahiro

- Indian Philosophy and Buddhist Studies // Center for Evolving Humanities
- Graduate School of Humanities and Sociology
- Tokyo University

Prof. Masahiro Shimoda will chair the panel History through Indic Texts.

3.19 Tomabechi, Toru

- International Institute for Digital Humanities, Tokyo
- toru.tomabechi@nifty.com

TEI Markup of Abhayākaragupta's Āmnāyamañjarī: An Attempt to Create an "Open Research Note" for the Study of Late Indian Buddhism

A text markup project may be conceived and designed in two different directions. The one direction is to create a sizable repository of multiple texts with basic (mainly structural) markup. The other is a type of "deep markup" project which aims at accumulating knowledge related to a particular work using e-text as information container. In this paper, we will report on the concept, objectives, and current status of our text markup project to create an "e-text-cum-philological-database" through deep TEI markup.

Since 2010, the "Vikramaśīla Project" (VP), funded with a JSPS grant and led by Prof. Taiken Kyuma (Mie University), has been putting together efforts of specialists in late Indian Buddhism to shed light on the relationship between Tantric and Non-Tantric Buddhist doctrines by laying focus on the works of authors associated with the Vikramaśīla monastery. One of those authors of special importance for the VP is Abhayākaragupta (11-12th c.), whose works are known for their richness in information. His magnum opus, the Āmnāyamañjarī (ĀM), is a commentary on the Sampuṭodbhavatantra and may be qualified "encyclopedic" in many respects. The VP has chosen the ĀM as platform to create an electronic research note on late Indian Buddhism. The research note has been designed as a deep-marked up TEI document, which is to be not only shared among the project members, but also made publicly available online. For our purposes, the ĀM is an excellent choice as material for several reasons. The text contains a large number of quotations from wide range of textual sources. In addition to explicit quotations, Abhayākaragupta also adopts, rather freely and often silently, many passages from works by his forerunners. Furthermore, he frequently refers to other texts of his own composition. Such references to external sources make the ĀM a very rich repository of historico-philological information. Reflecting Abhayākaragupta's wide and deep erudition, the contents of the text also cover a wide range of subjects well beyond the boundary of a mere Tantric commentary. This latter character of the ĀM renders the text itself a vast subject- and terminology-inventory quite useful for the study of late Indian Buddhism.

By encoding the philological, doctrinal and lexical information contained in the ĀM, we have been trying to explore the potential of the TEI-compliant textual markup for the study of Indian Buddhism and to, at the same time, probe into both the strength and the limitation of the current TEI guidelines. In the course of the collaborative work among the project members, we encountered a number of questions to be addressed: What is the best way to structuralize the document? – Which TEI element is appropriate for a particular information in the text? – How the collaboration should be organized? – How the result is to be shared? – and so forth. This paper is a (pre-)interim

report of our attempt to create an electronic research note which is "open" in several senses – by open collaboration, following open standards, for open access and designed as a platform of open-end accumulation of knowledge. Critical comments, advices, and other inputs from the experienced participants of the Workshop are sincerely welcome.

3.20 Wujastyk, Dominik

- wujastyk@gmail.com
- Singhmar Chair in Classical Indian Society and Polity, Department of History and Classics, University of Alberta, Canada

What do users want from SARIT in future?

This presentation will discuss the place of e-texts in contemporary scholarship, and the possibilities that e-texts may afford our successors. It will also interrogate the place of SARIT in this scheme, and why encoding standards are of central importance.

The SARIT library has matured substantially in the last five years. There are more texts, more highly developed guidelines for how texts are encoded, more sophisticated search facilities and a reimplementation of the whole platform on a more robust software basis. The direction of these developments has been driven principally by the views and requirements of the SARIT team themselves. This is good, and as it should be. Yet there are some consequences. SARIT has not yet emerged as the go-to service for Indic e-texts. That position is still held by the GRETIL service, in spite of its technical and qualitative inferiority. The SARIT project has not been as successful as it could be in changing the public perception amongst Indologists regarding the virtues of textual integrity, reliability, version control, and responsibility. Most scholars are aware of the value of a well-edited printed edition of an Indic text, but that same attitude has not yet become widespread where electronic texts are concerned. Furthermore, SARIT has not been successful in broadcasting updates to either its growing content or its developing technical features. SARIT has not consulted widely about the features that the general indological public most want from a library of electronic texts.

This presentation will address the strengths of the SARIT library as it exists in spring 2017, and present the results of a public user-consultation about directions for future development.

3.21 Zotter, Christof and Manik Bajracharya

- christof.zotter@adw.uni-heidelberg.de
- Documents on the History of Religion and Law of Pre-modern Nepal, Heidelberg Academy of Science and Humanities

See Bajracharya, Manik and Christof Zotter for the abstract.

3.22 Software demonstrations

This module is intended for the presentation of practical aspects mentioned in the talks.

1. Yigal Bronner: Hands-on presentation of the Panditproject.org database
2. Amba Kulkarni: Demonstration of Saṁsādhanī: A platform of Sanskrit Computational Tools
3. Stefan Baums: Research Environment for Ancient Documents (READ)
4. Peter M. Scharf: Presentation of the Sanskrit Library meter identification tool and the integrated dictionary interface
5. Patrick McAllister: Presentation of SARIT's workflow for editing and displaying Indic XML documents

4 Links

- https://asiabeyondboundaries.org/about/: Homepage 'Beyond Boundaries: Religion, Region, Language and the State' (see Dániel Balogh)
- http://east.uni-hd.de: Epistemology and Argumentation in South Asia and Tibet (see Birgit Kellner)
- http://gandhari.org: Gāndhārī Language and Literature (see Stefan Baums)
- http://github.com/chchch/upama: *upama*, library to compare Sanskrit TEI XML files and generate an apparatus (see Charles Li)
- http://historyandclassics.ualberta.ca/: Department of History and Classics, University of Alberta, Canada (see Dominik Wujastyk)
- http://panditproject.org: Prosopographical Database for Indic texts (see Yigal Bronner)
- http://saktumiva.org/wiki:dravyasamuddesa:start: Edition "The Dravyasamuddeśa of Bhartṛhari" (see Charles Li)
- http://sanskritlibrary.org: The Sanskrit Library (see Peter M. Scharf)
- http://sarit.indology.info: SARIT — Search and Retrieval of Indic Texts
- http://www.haw.uni-heidelberg.de/forschung/forschungsstellen/nepal/index.de.html: Documents on the History of Religion and Law of Pre-modern Nepal (see Bajracharya, Manik and Christof Zotter)
- http://www.haw.uni-heidelberg.de/forschung/forschungsstellen/nepal/publ_docs.en.html: Published documents on the history of religion and law of premodern Nepal (see Bajracharya, Manik and Christof Zotter)
- http://www.ikga.oeaw.ac.at/Mainpage: Homepage of the Institute for the Cultural and Intellectual History of Asia, Austrian Academy of Sciences
- http://www.mitpressjournals.org/doi/pdf/10.1162/COLI_a_00239: Deep learning (see Oliver Hellwig)

- http://www.r2d3.us/visual-intro-to-machine-learning-part-1/: Introduction to machine learning (see Oliver Hellwig)
- http://www.stockholmuniversitypress.se/site/books/10.16993/baj/: P. Robinson, "The Digital Revolution in Scholarly Editing." (see Philipp Maas)
- https://gandhari.org/blog/?p=251: Blog post about the "Research Environment for Ancient Documents" (see Stefan Baums)
- https://github.com/readsoftware/read: Software repository for the "Research Environment for Ancient Documents" (see Stefan Baums)
- https://projecteuclid.org/download/pdfview_1/euclid.ss/1089808279: Introduction to graphical models (see Oliver Hellwig)